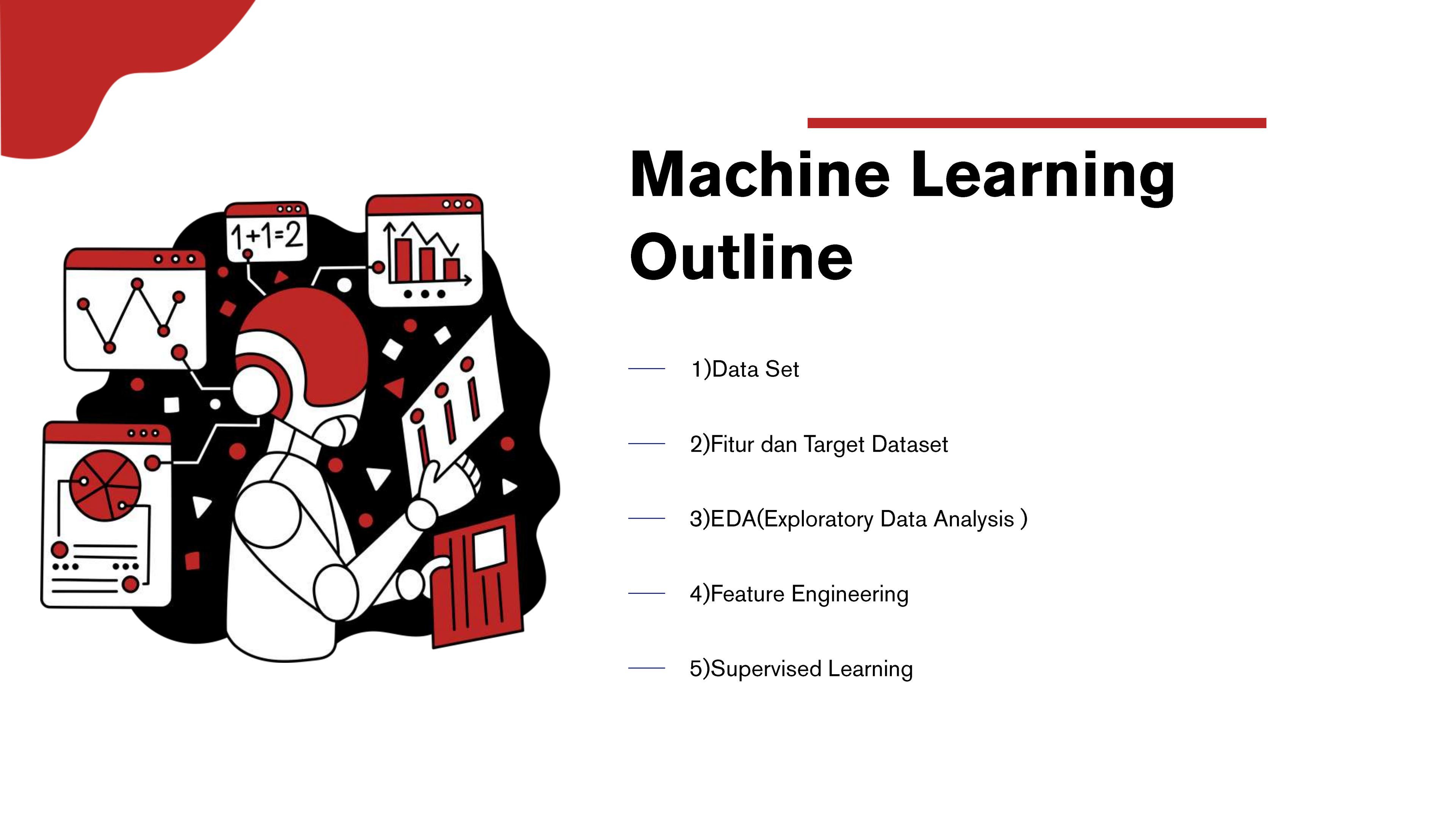




Company Clustering

Presented by: Team 1 <https://www.kaggle.com/code/artamevia/company-clustering>



Machine Learning Outline



- 1) Data Set
- 2) Fitur dan Target Dataset
- 3) EDA(Exploratory Data Analysis)
- 4) Feature Engineering
- 5) Supervised Learning

1

DataSet

Company Clustering (pengelompokan perusahaan)

Statistik Umum:

Tujuan utama dari analisis ini adalah untuk mengelompokkan perusahaan ke dalam beberapa klaster berdasarkan kesamaan karakteristik mereka.

Jumlah sampel: 167 perusahaan.

employee_turnover: Rata-rata 38.27, dengan variasi cukup besar (standar deviasi 40.33), menunjukkan perbedaan signifikan dalam tingkat turnover karyawan antar perusahaan.

revenue_growth: Rata-rata pertumbuhan pendapatan adalah 41.11%, tetapi variasinya cukup besar (standar deviasi 27.41).

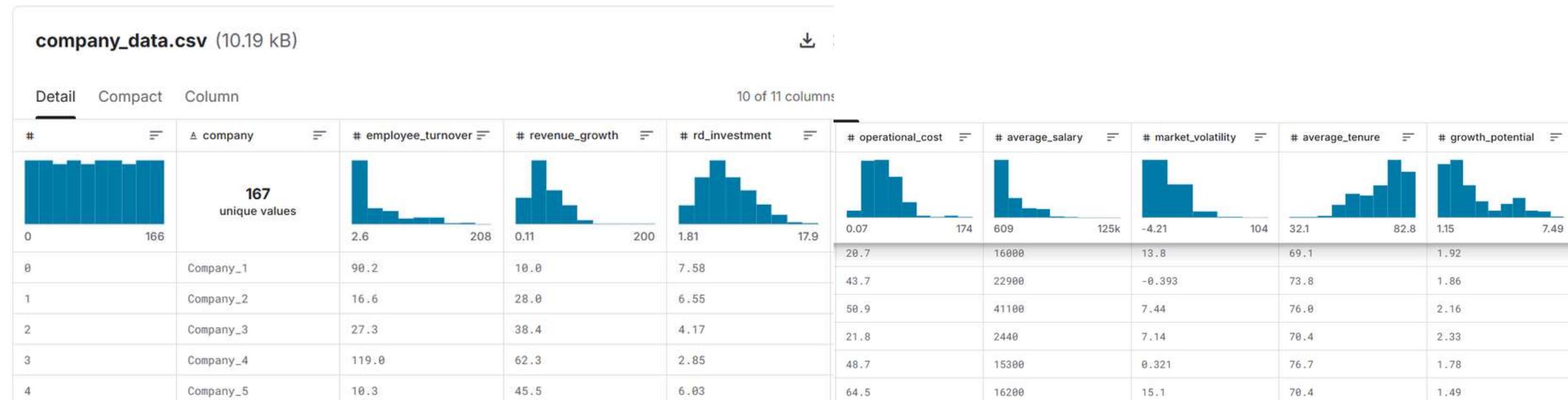
rd_investment: Rata-rata investasi R&D adalah 6.82%, dengan variasi yang relatif lebih kecil.

average_salary: Rata-rata gaji sangat bervariasi (standar deviasi 19278.07) dengan rata-rata 17144.69.

net_profit: Laba bersih juga menunjukkan variasi yang signifikan antar perusahaan.

<https://www.kaggle.com/code/artamevia/company-clustering>

Input Data



2

Fitur & Target DataSet

Company Clustering

- 1.Unnamed: 0: Kolom ini sepertinya adalah indeks baris, tidak berhubungan langsung dengan data.
- 2.Company: Menampilkan nama perusahaan fiktif, sebagai identifikasi unik setiap perusahaan.
- 3.employee_turnover: Mengukur turnover karyawan per 1000 karyawan. Nilai tinggi menandakan tingkat pergantian karyawan yang lebih sering.
- 4.revenue_growth (pertumbuhan pendapatan): Menunjukkan pertumbuhan pendapatan perusahaan dalam persentase. Nilai positif menandakan pertumbuhan, sementara nilai negatif bisa menandakan penurunan.
- 5.rd_investment (investasi rd): Mengindikasikan berapa persentase pengeluaran perusahaan untuk penelitian dan pengembangan. Semakin tinggi, semakin banyak investasi dalam inovasi.
- 6.operational_cost (biaya operasional): Menunjukkan biaya operasional sebagai persentase dari total pengeluaran. Nilai rendah menandakan efisiensi operasional yang lebih baik.
- 7.average_salary (rata-rata pendapatan/gaji): Rata-rata gaji karyawan perusahaan. Nilai tinggi bisa menunjukkan skala ekonomi yang lebih besar atau strategi perekutan yang berfokus pada kualitas.
- 8.market_volatility (volatilitas pasar) : Mengukur tingkat volatilitas pasar tempat perusahaan beroperasi. Nilai tinggi menandakan lingkungan bisnis yang lebih tidak terprediksi.
- 9.average_tenure (rata-rata kontrak karyawan): Rata-rata lama karyawan bekerja di perusahaan dalam tahun. Nilai tinggi bisa menandakan loyalitas dan kepuasan kerja yang lebih besar.
- 10.growth_potential (potensi pertumbuhan): Skala untuk menilai potensi pertumbuhan perusahaan dari 1 hingga 10. Angka yang lebih tinggi menandakan potensi yang lebih besar.
- 11.net_profit (laba bersih): Menunjukkan laba bersih perusahaan. Angka ini penting untuk menilai kesehatan finansial dan kesuksesan perusahaan.

<https://www.kaggle.com/code/artamevia/company-clustering>

Unnamed: 0	company	employee_turnover	revenue_growth	rd_investment	operational_cost	average_salary	market_volatility	average_tenure	growth_potential	net_profit
0	0	Company_1	90.2	10.0	7.58	44.9	1610	9.44	56.2	5.82
1	1	Company_2	16.6	28.0	6.55	48.6	9930	4.49	76.3	1.65
2	2	Company_3	27.3	38.4	4.17	31.4	12900	16.10	76.5	2.89
3	3	Company_4	119.0	62.3	2.85	42.9	5900	22.40	60.1	6.16
4	4	Company_5	10.3	45.5	6.03	58.9	19100	1.44	76.8	2.13

Output

3

EDA(Exploratory Data Analysis)

proses analisis awal untuk memahami karakteristik dataset sebelum melakukan pemodelan lebih lanjut.

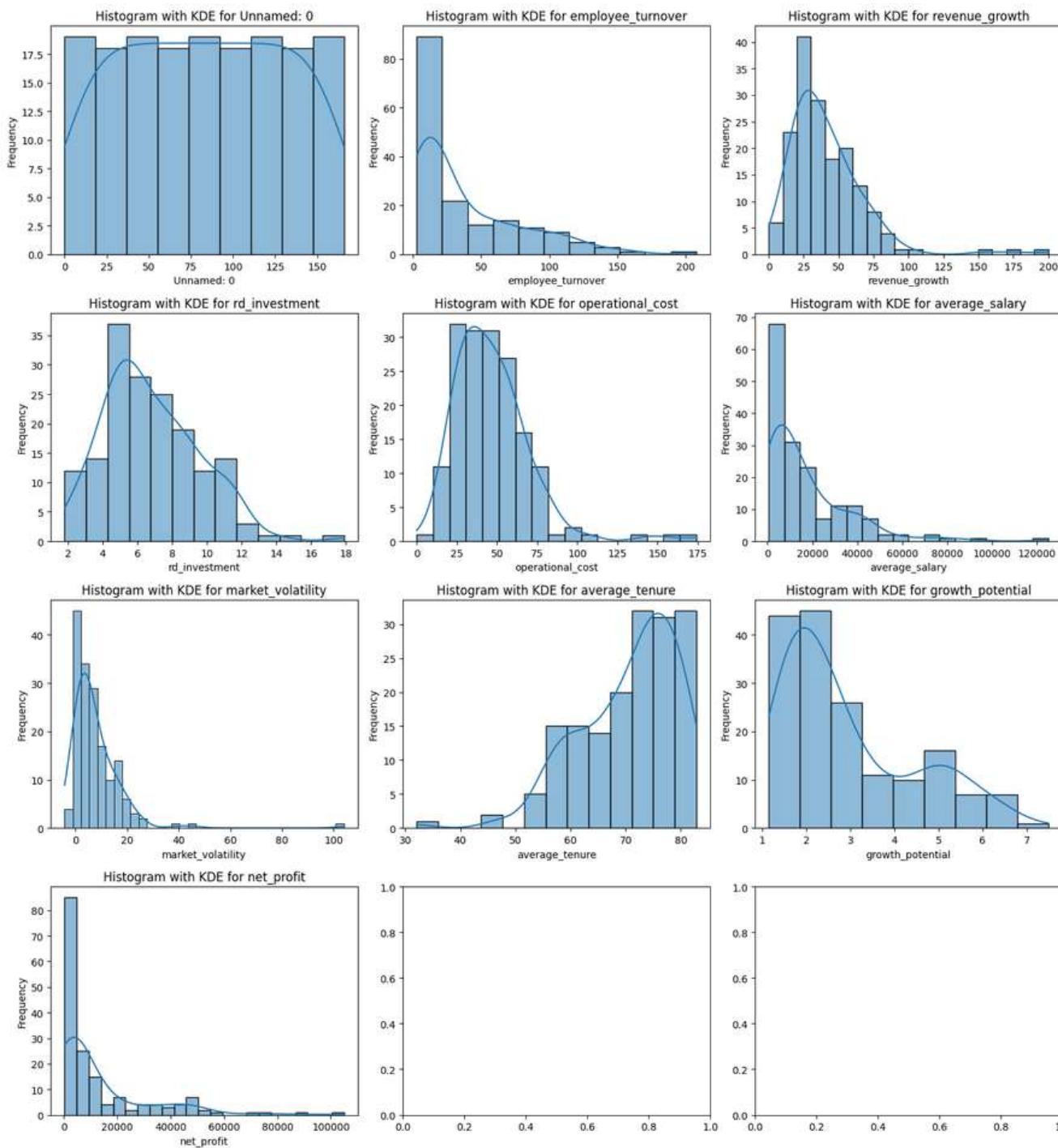
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Unnamed: 0        167 non-null    int64  
 1   company          167 non-null    object  
 2   employee_turnover 167 non-null    float64 
 3   revenue_growth   167 non-null    float64 
 4   rd_investment    167 non-null    float64 
 5   operational_cost 167 non-null    float64 
 6   average_salary   167 non-null    int64  
 7   market_volatility 167 non-null    float64 
 8   average_tenure   167 non-null    float64 
 9   growth_potential 167 non-null    float64 
 10  net_profit       167 non-null    int64  
dtypes: float64(7), int64(3), object(1)
memory usage: 14.5+ KB
```

Informasi Data

3

EDA(Exploratory Data Analysis)

Grafik hasil data EDA



Duplicate Handling

Outlier Handling

Missing Value Handling

Data Encoding

Feature Scalling

4

Feature Engineering

Feature Engineering adalah alat penting untuk meningkatkan performa model

	Total Missing	Percentage Missing (%)
company	0	0.0
employee_turnover	0	0.0
revenue_growth	0	0.0
rd_investment	0	0.0
operational_cost	0	0.0
average_salary	0	0.0
market_volatility	0	0.0
average_tenure	0	0.0
growth_potential	0	0.0
net_profit	0	0.0

Missing value

Duplicate Handling

```
[ ] len(df) #mengecek panjang baris dari DataFrame
→ 167

▶ len(df) - len(df.drop_duplicates()) #mengecek selisih antara panjang DataFrame (len(df)) dan menguranginya dengan jumlah baris yang duplikat
→ 0
```

Pada kode (`len(df) - len(df.drop_duplicates())`), kita menghitung selisih antara panjang DataFrame (`len(df)`) dan menguranginya dengan jumlah baris yang duplikat. Ini menghitung semua duplikat di DataFrame.

```
[ ] len(df.drop_duplicates()) #mengecek panjang baris yang duplikat
→ 167
```

kita juga bisa melihat baris yang duplikat

```
[ ] len(df.drop_duplicates()) / len(df) #jika output = 1.0 maka tidak ada duplikat
→ 1.0
```

Drop Duplikat

Feature Engineering

Splitting Data

metode membagi data menjadi dua bagian atau lebih yang membentuk subhimpunan data. Umumnya, data splitting memisahkan dua bagian, bagian pertama digunakan untuk mengevaluasi atau uji data dan data lainnya digunakan untuk melatih model.

Hasil Splitting Data:

...	Train size :	Unnamed: 0	employee_turnover	revenue_growth	rd_investment	\
0	0	90.2	10.00	7.58		
82	82	10.8	66.70	2.63		
59	59	74.7	29.50	5.22		
95	95	7.9	86.90	4.39		
60	60	3.9	22.10	10.30		
..	
17	17	111.0	23.80	4.10		
98	98	6.8	153.00	8.65		
66	66	208.0	15.30	6.91		
126	126	63.6	12.00	10.50		
109	109	47.0	9.58	5.25		

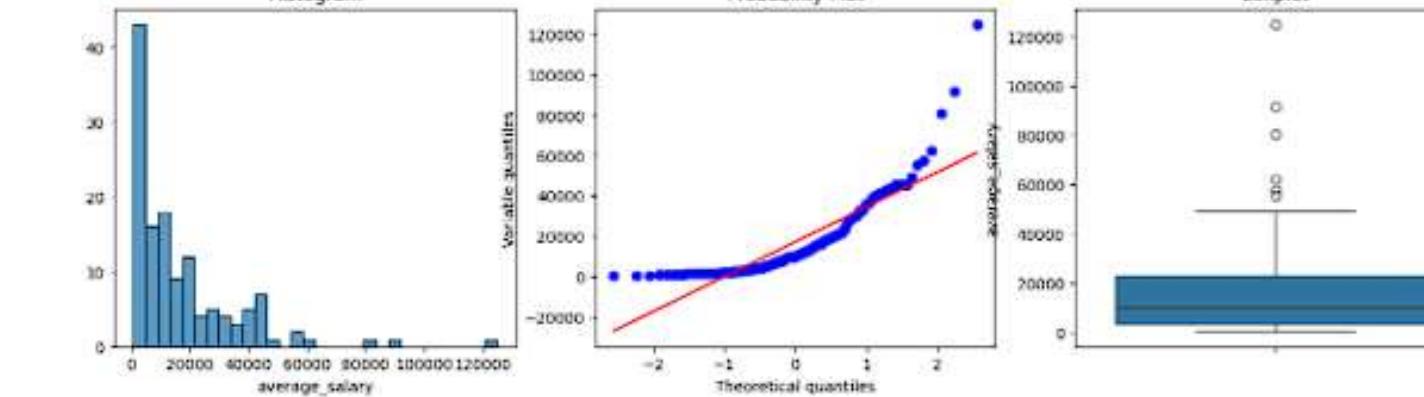
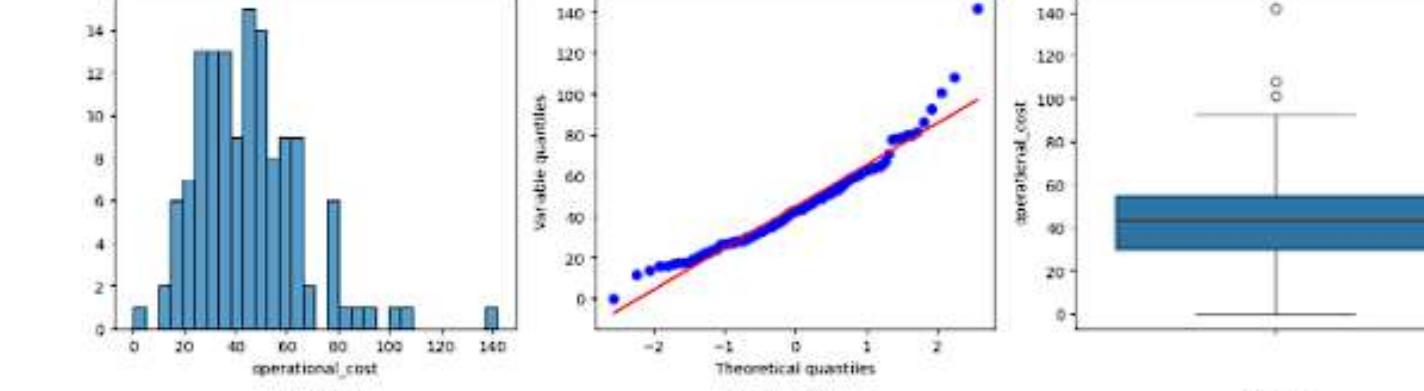
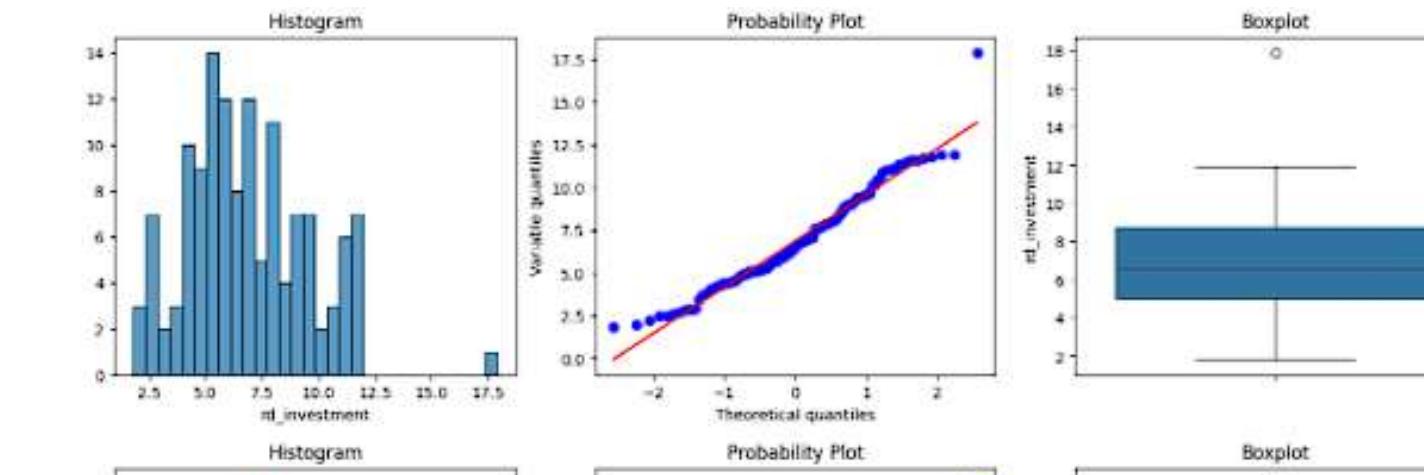
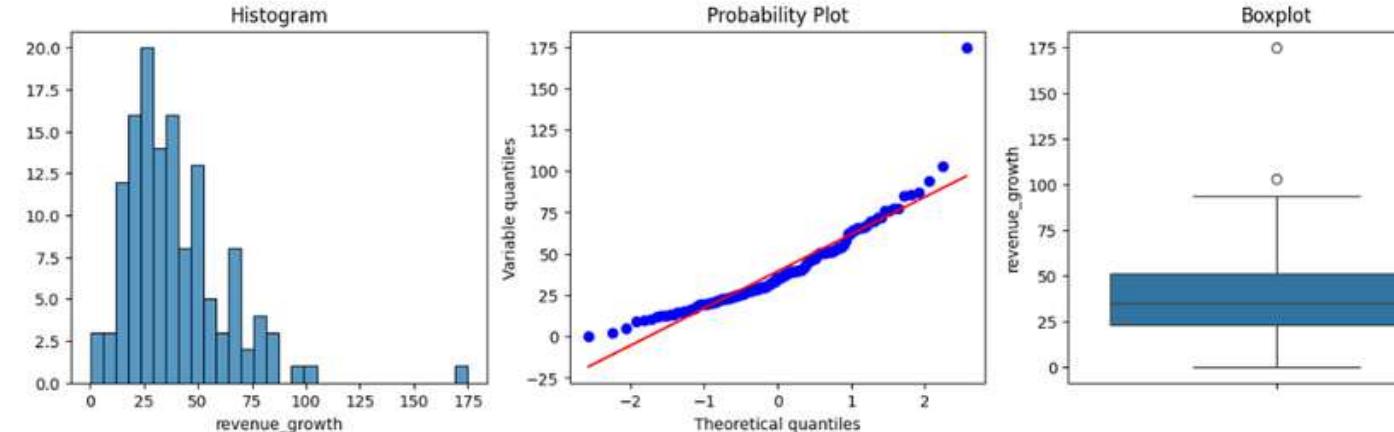
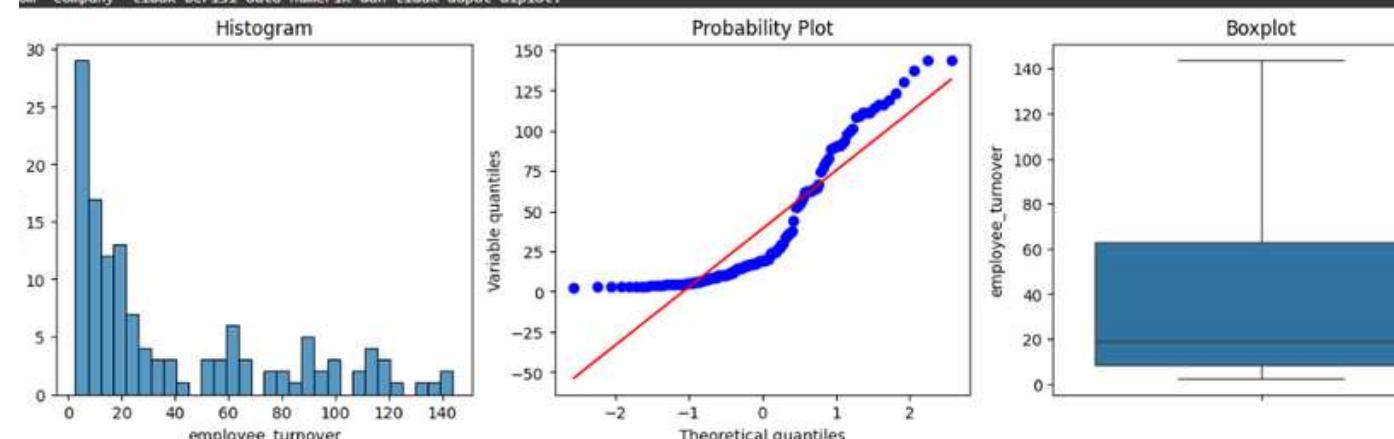
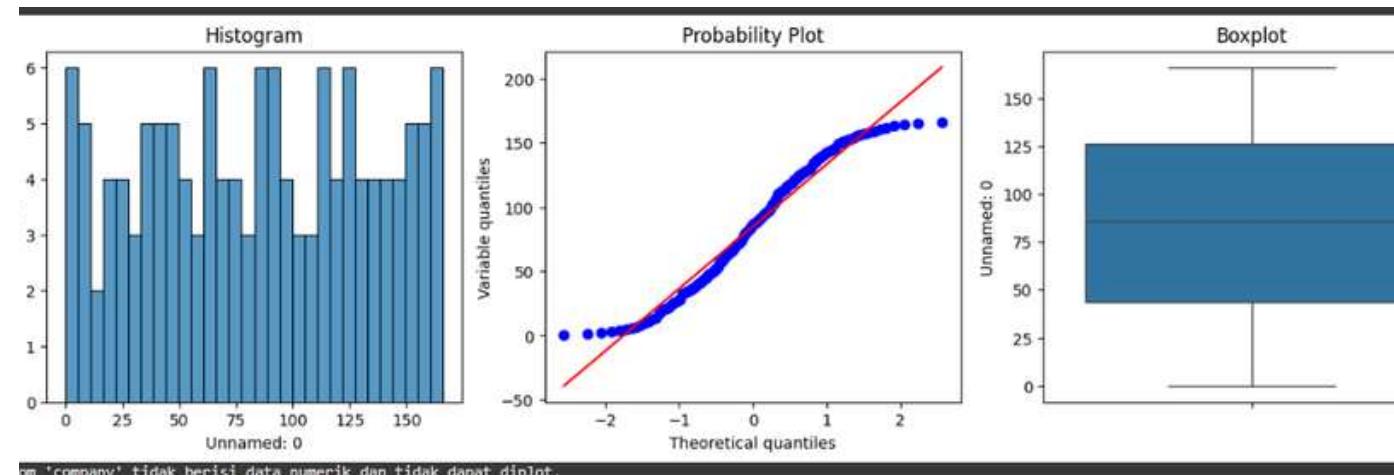
...	operational_cost	average_salary	market_volatility	average_tenure	\
0	44.9	1610	9.440	56.2	
82	30.4	75200	11.200	78.2	
59	45.9	3060	16.600	62.2	
95	71.0	21100	7.270	74.5	
60	30.7	28700	0.673	80.4	
..	
17	37.2	1820	0.885	61.8	
98	154.0	28300	3.830	80.3	
66	64.7	1500	5.450	32.1	
126	30.0	1350	2.610	64.6	
109	36.4	1990	15.100	68.3	
...					
19	3.20				
87	3.30				
29	1.63				
35	2.01				

Output is truncated. View as a [scrollable element](#) or open in a [text editor](#). Adjust cell output [settings](#)...

4

Feature Engineering - Outlier Handling

Feature Engineering adalah alat penting untuk meningkatkan performa model

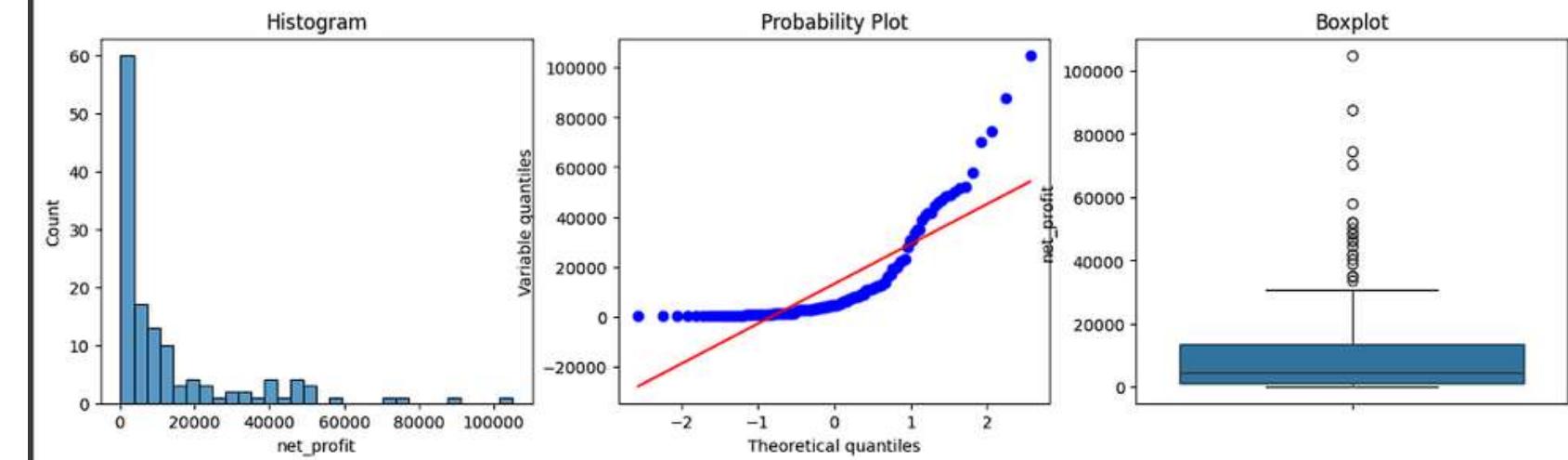
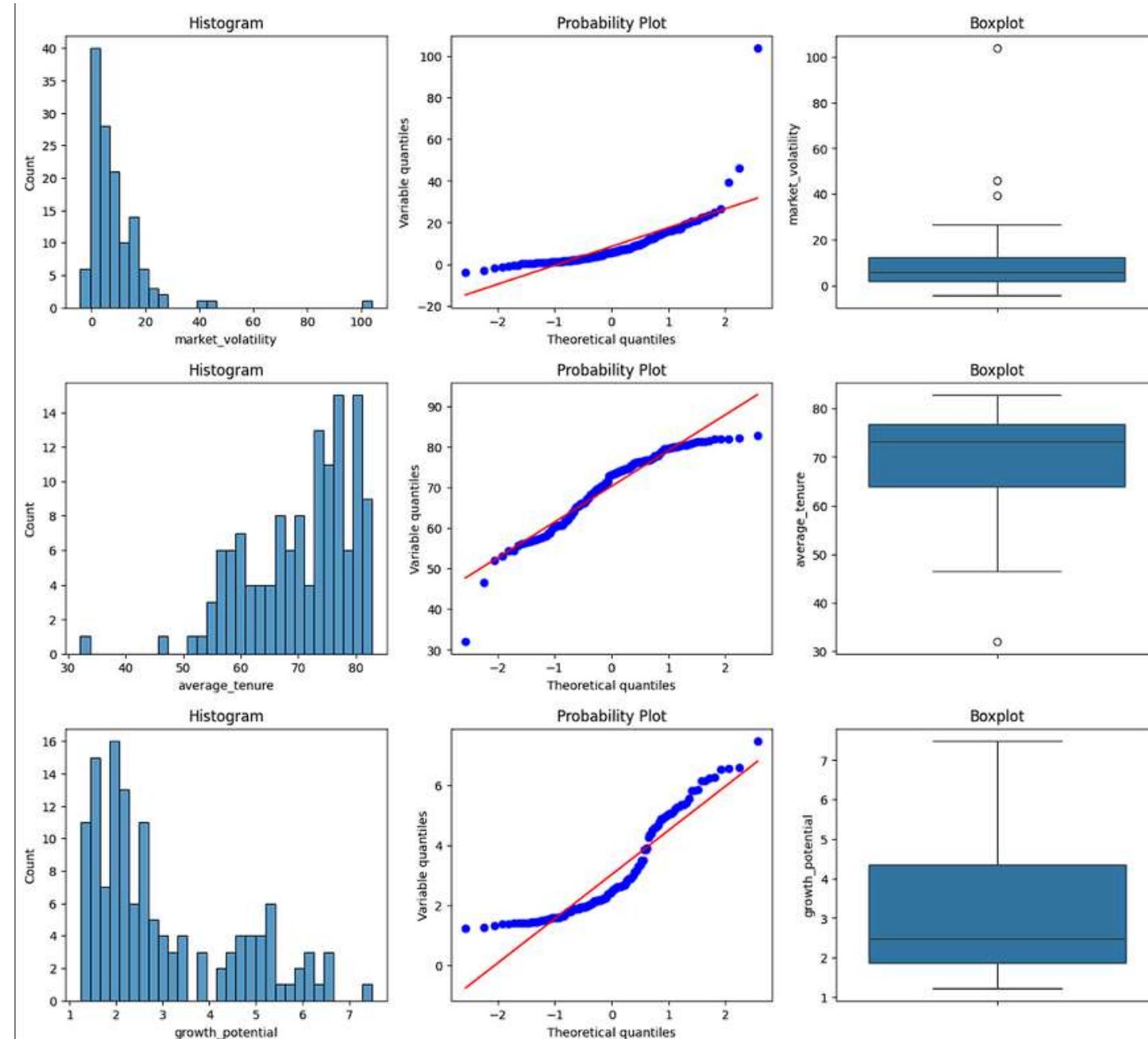


Outlier Handling Sebelum

4

Feature Engineering

Feature Engineering adalah alat penting untuk meningkatkan performa model

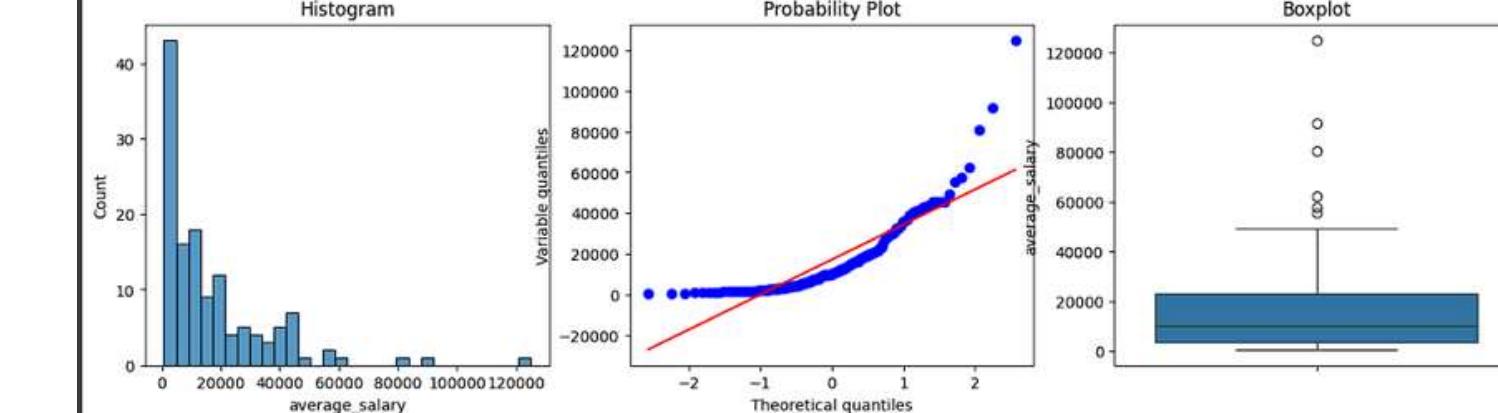
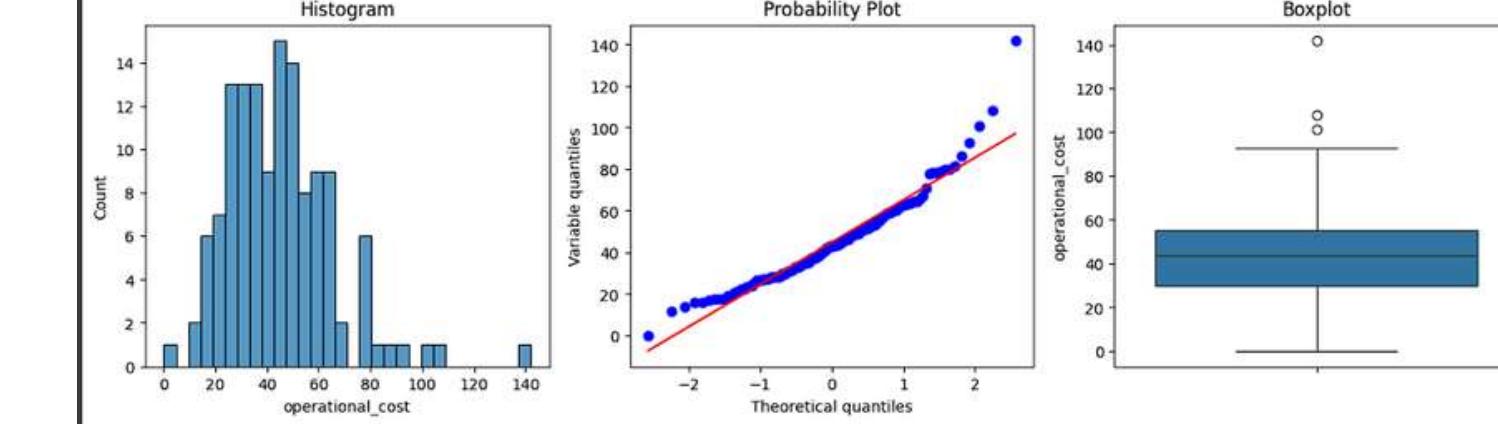
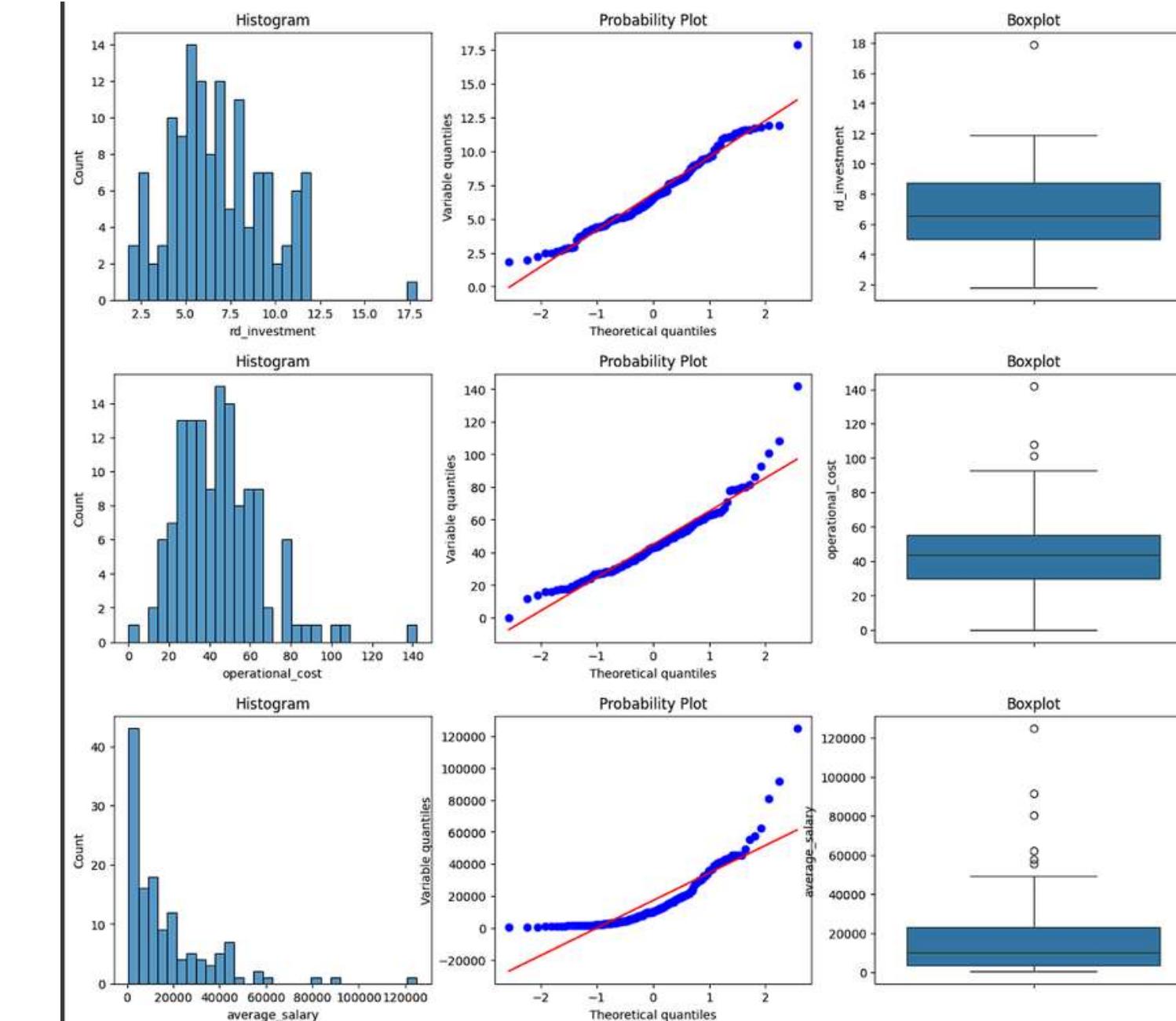
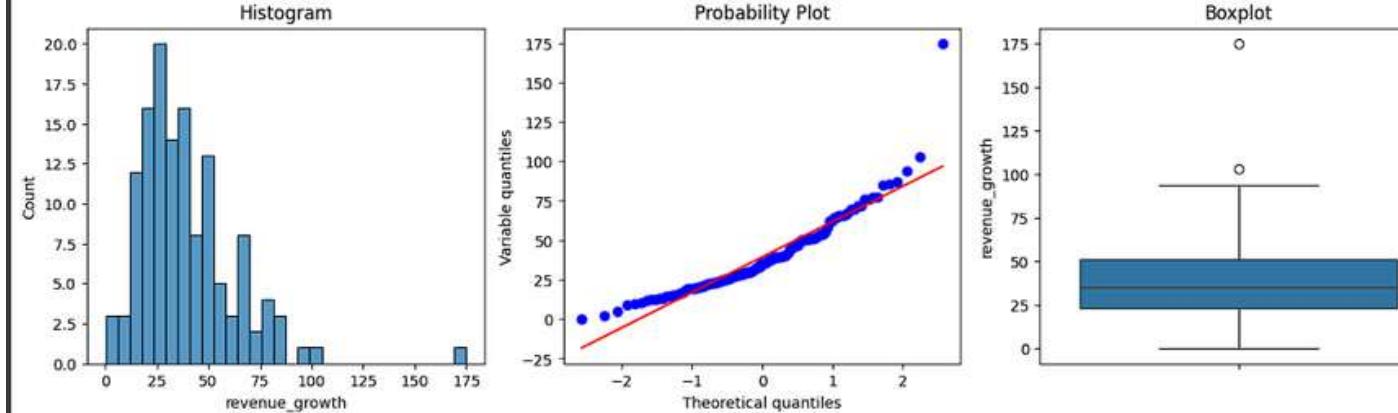
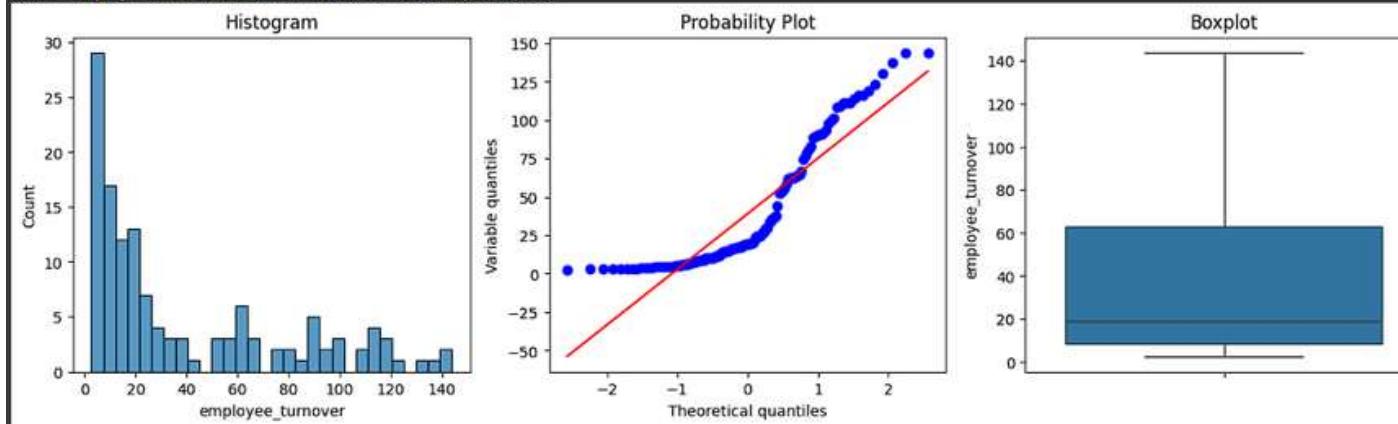
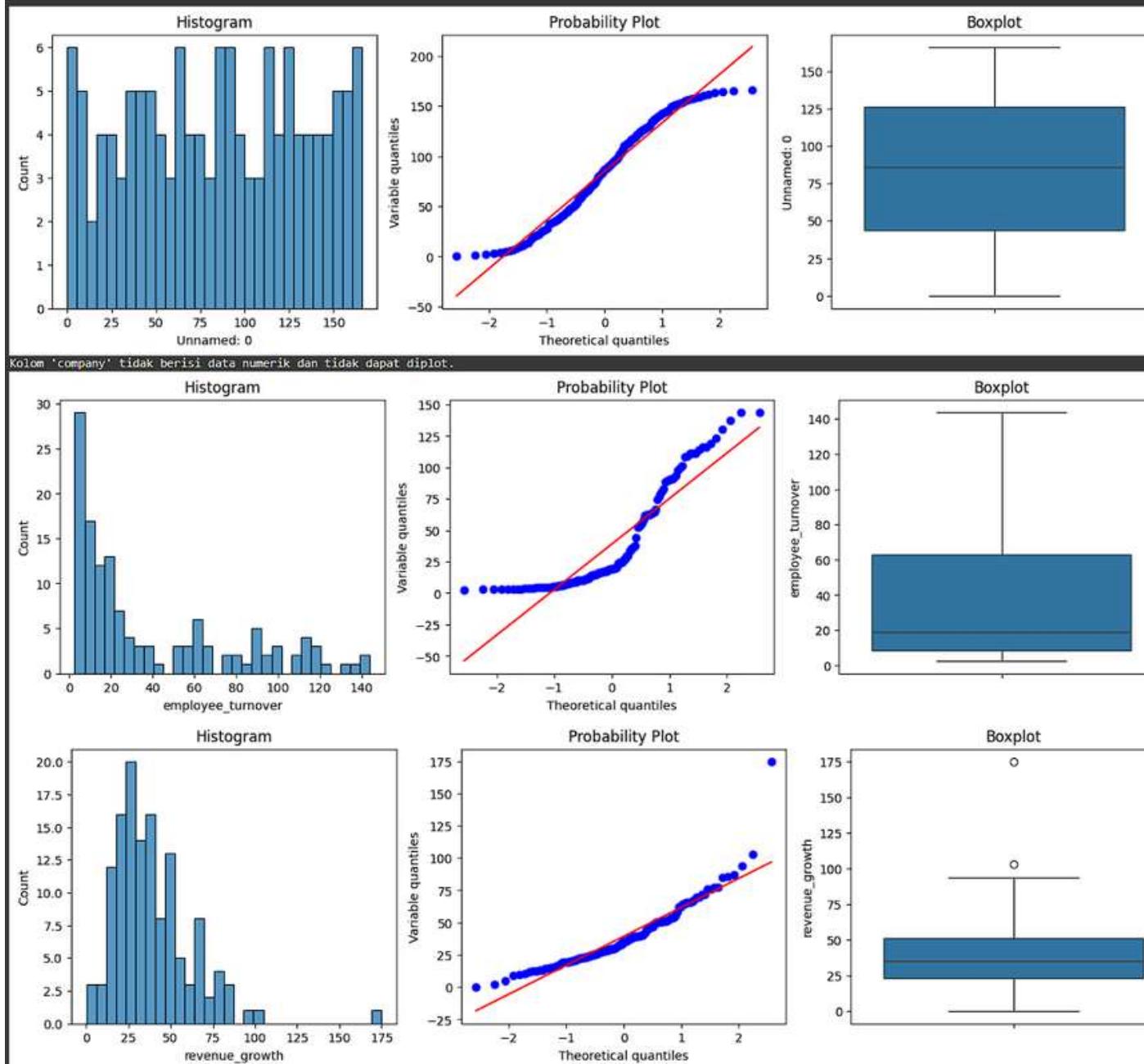


Outlier Handling Sebelum

4

Feature Engineering

Feature Engineering adalah alat penting untuk meningkatkan performa model

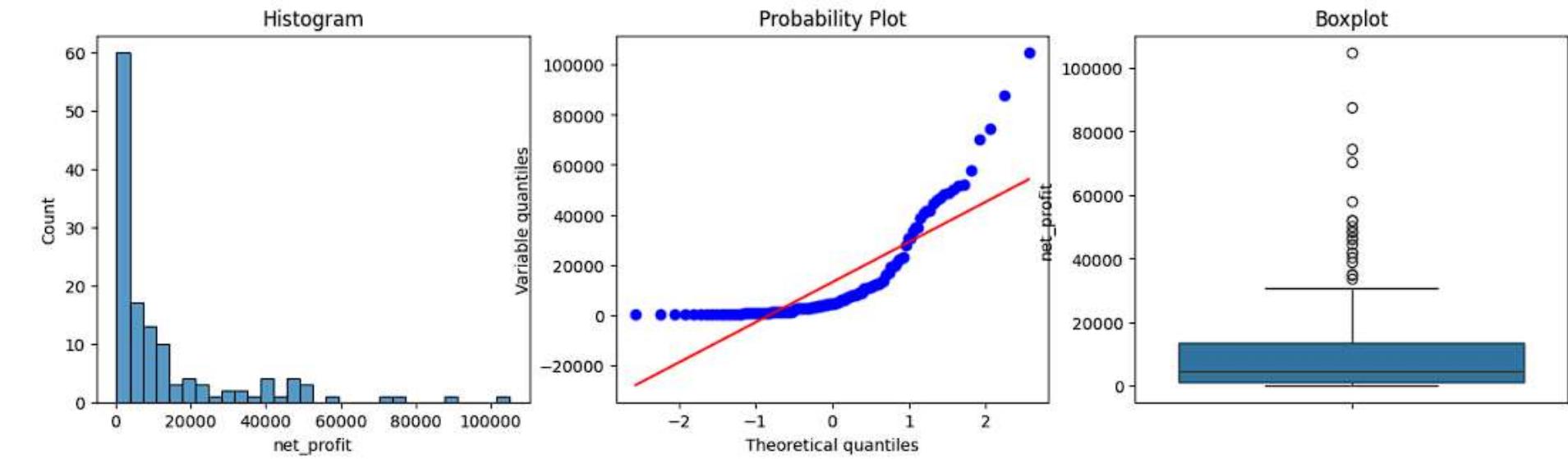
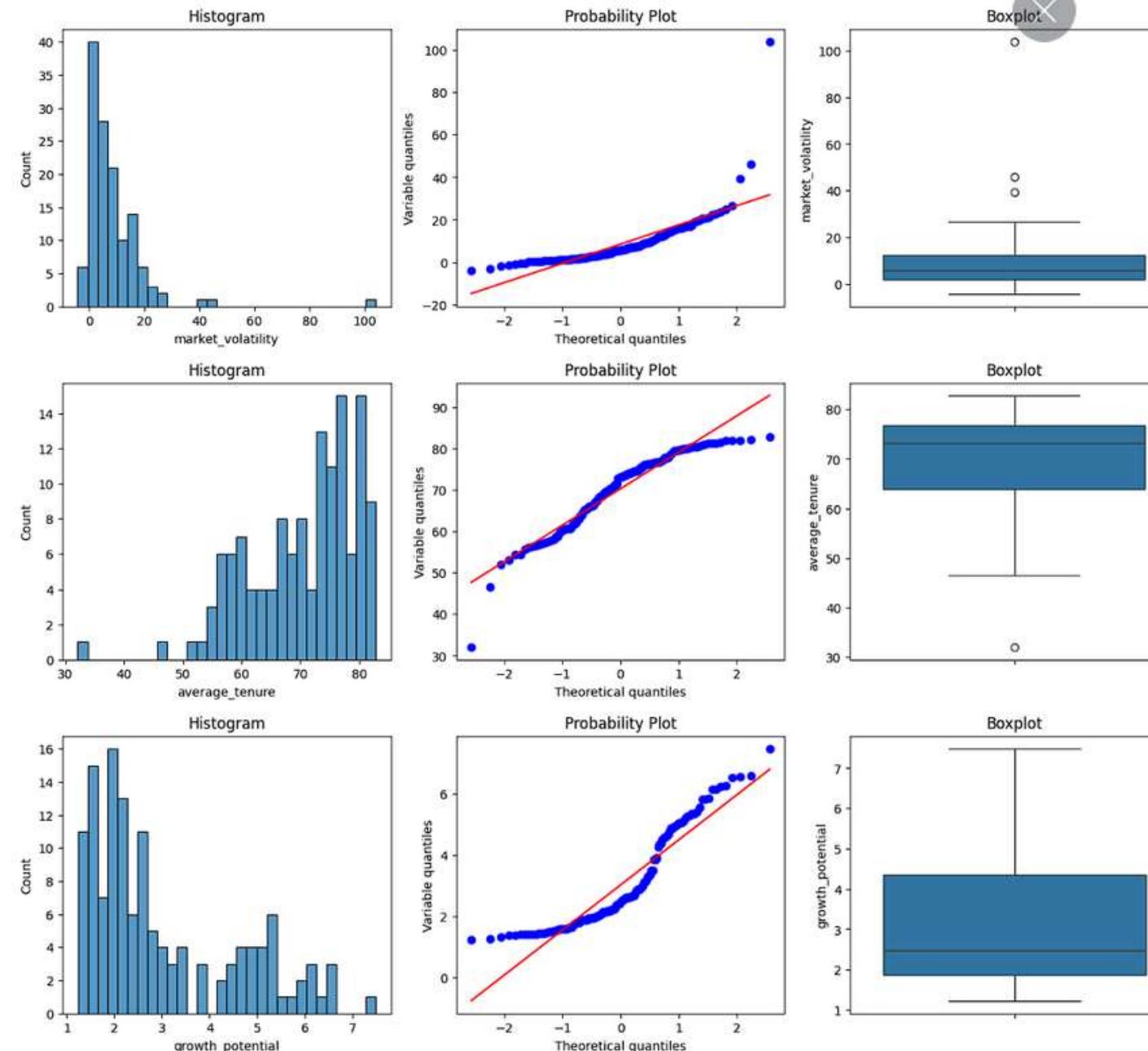


Outlier Handling Sesudah

4

Feature Engineering

Feature Engineering adalah alat penting untuk meningkatkan performa model

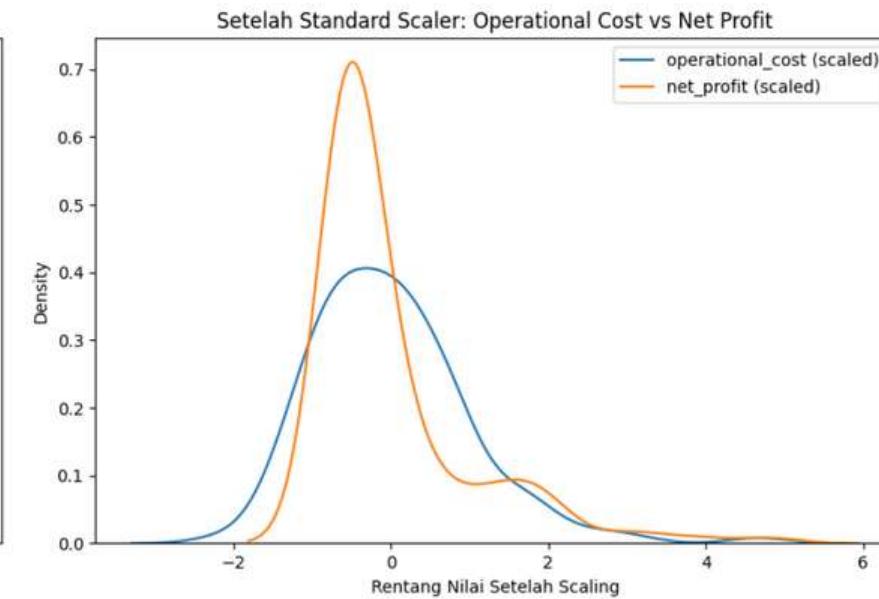
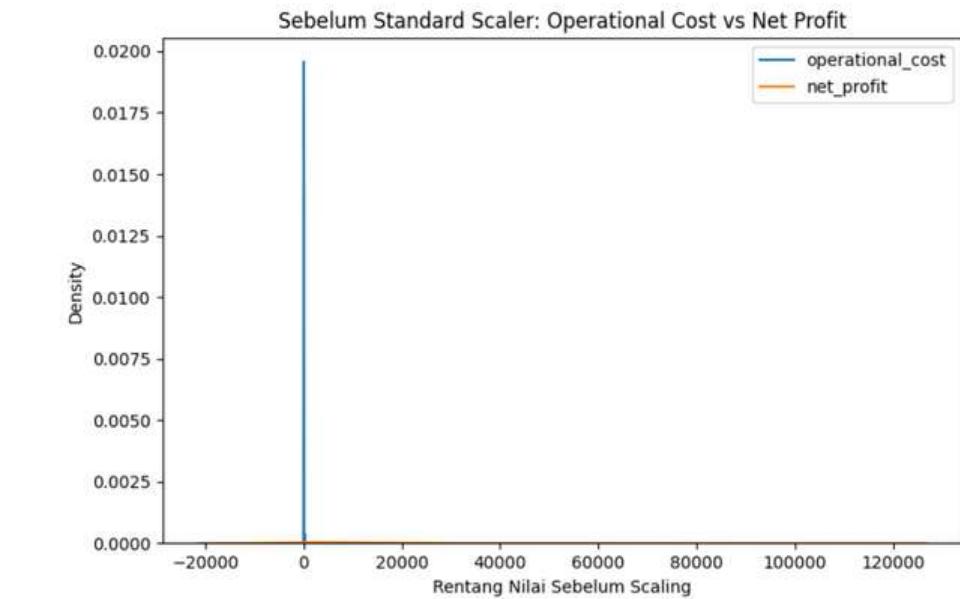
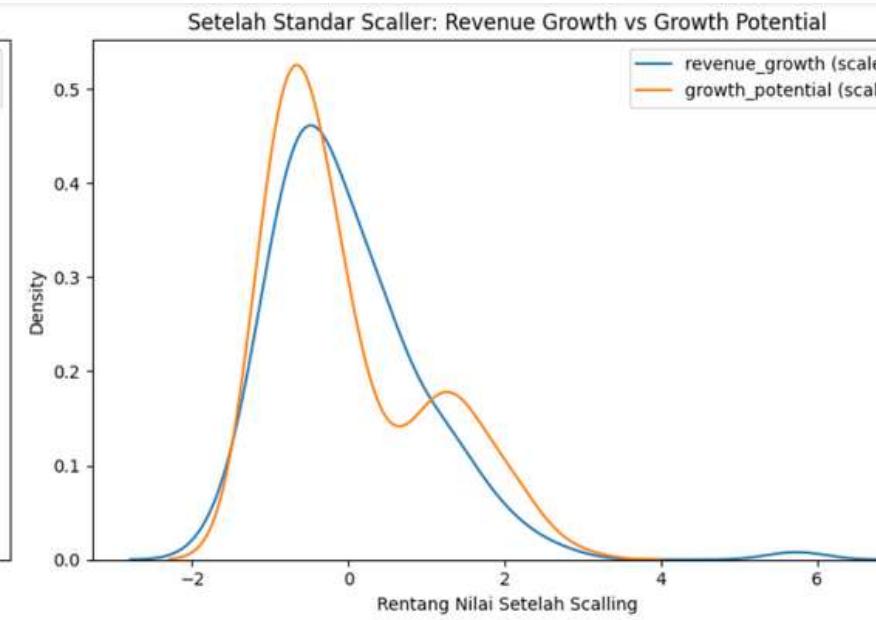
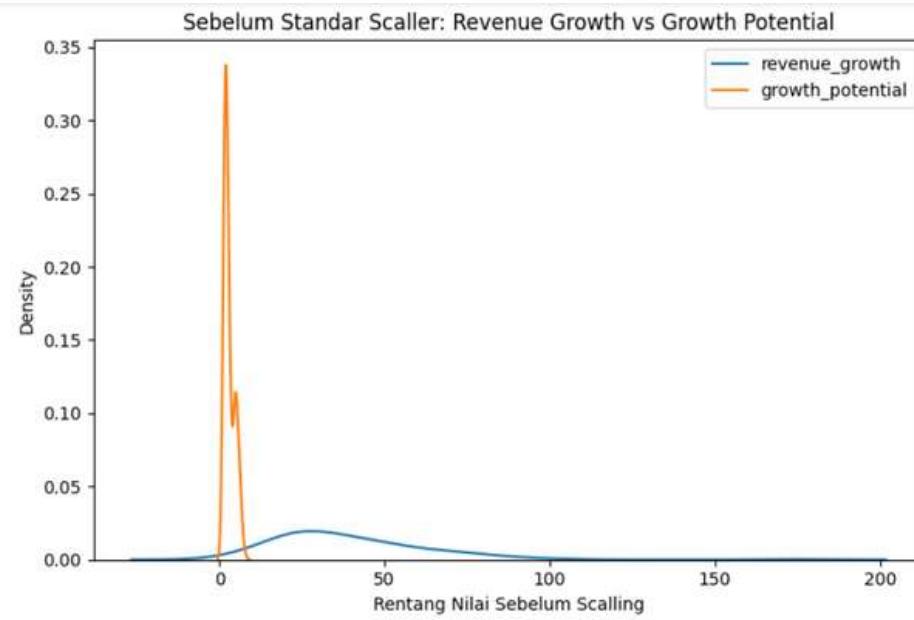


Outlier Handling Sesudah

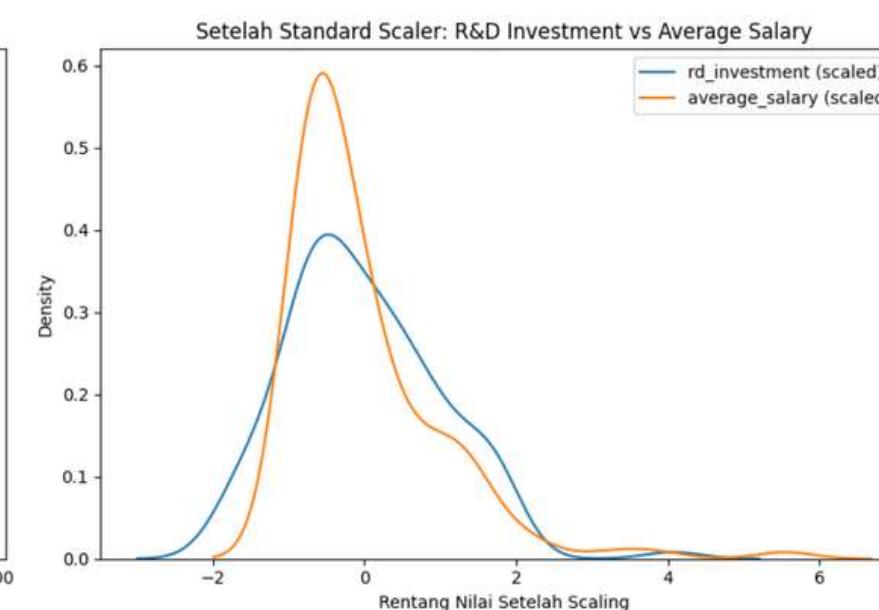
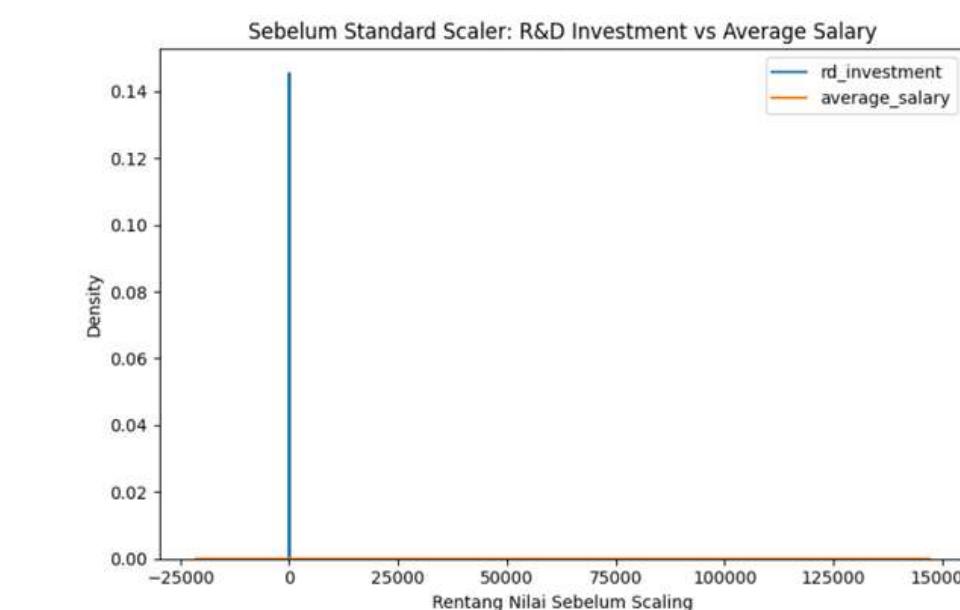
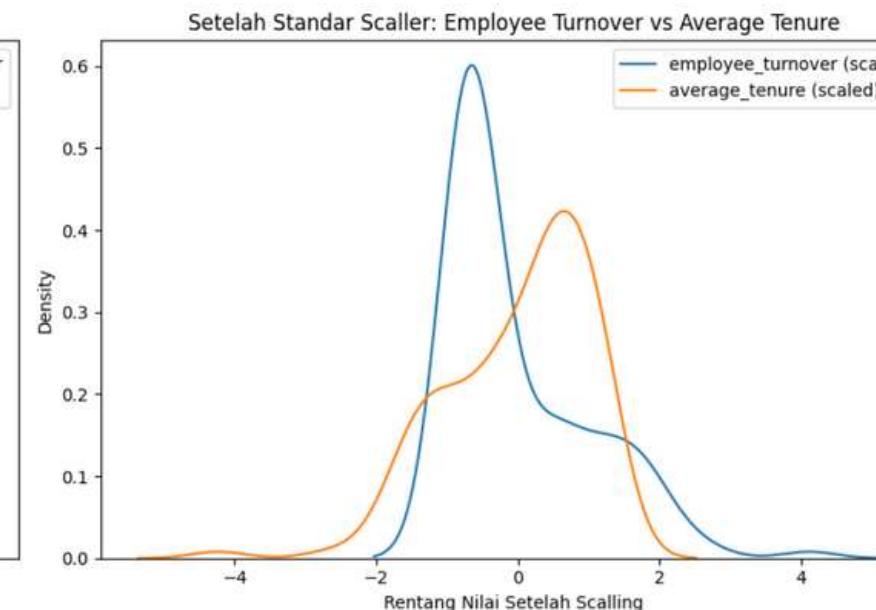
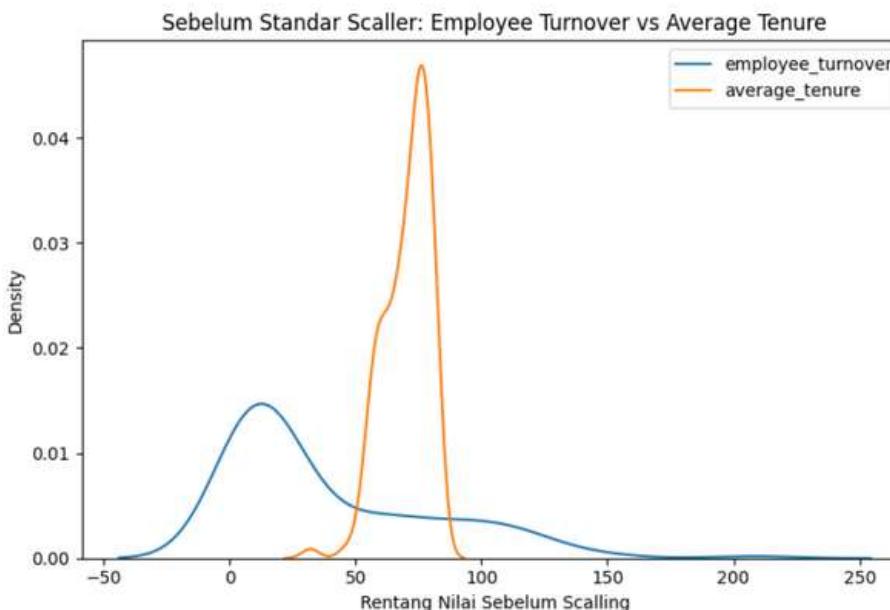
4

Feature Engineering

Standar Scalling



Revenue_Growth & Growth_potential



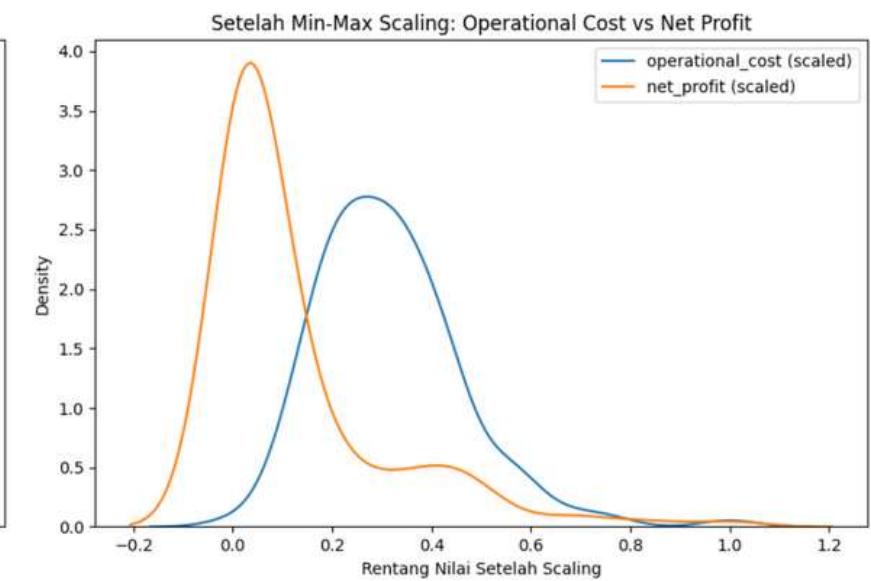
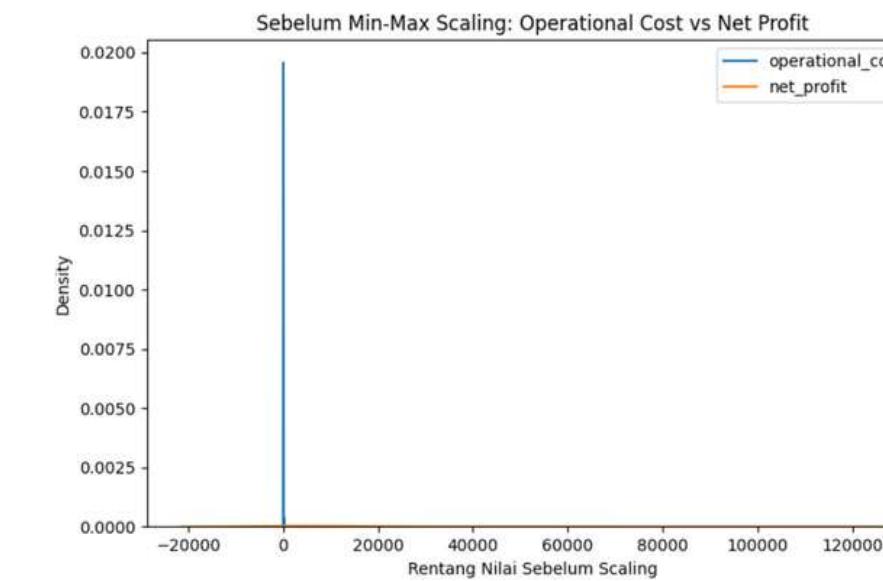
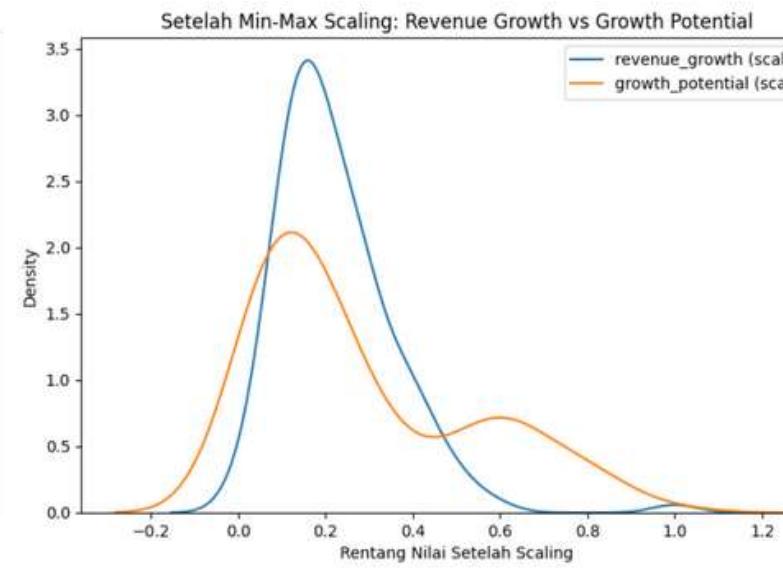
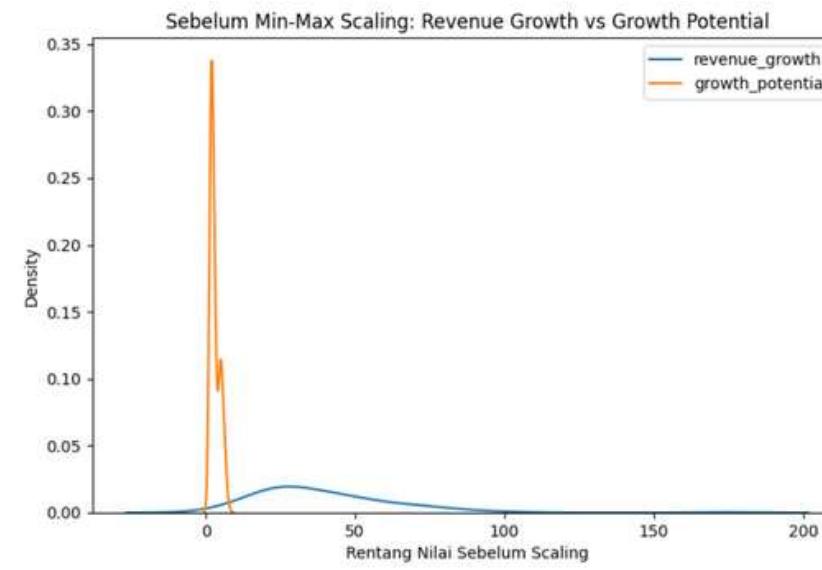
Employee_turnover & average_tenure

rd_invesment & Average Salary

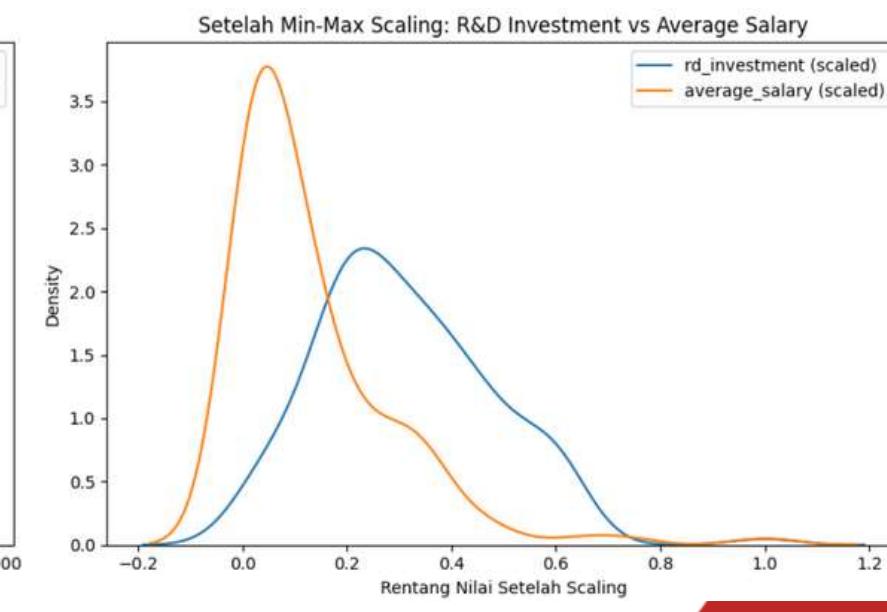
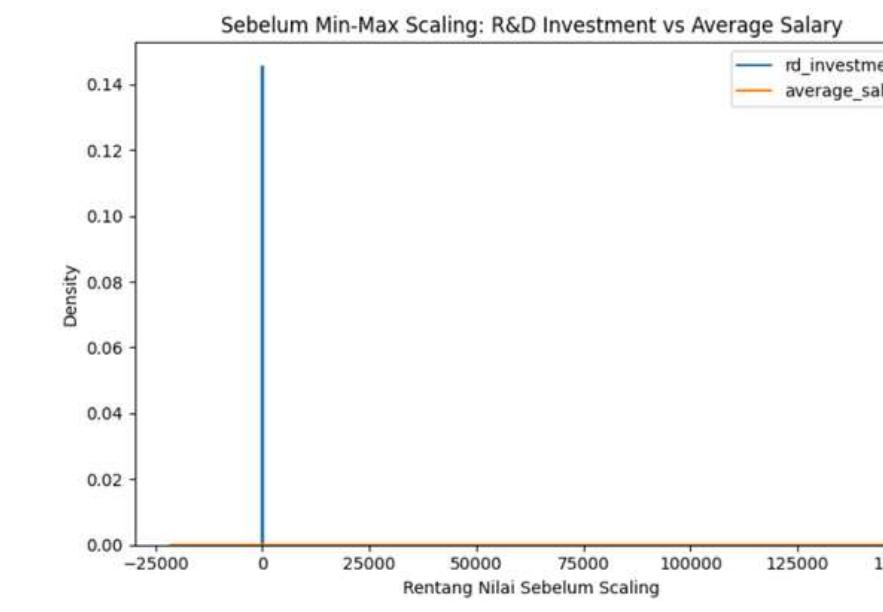
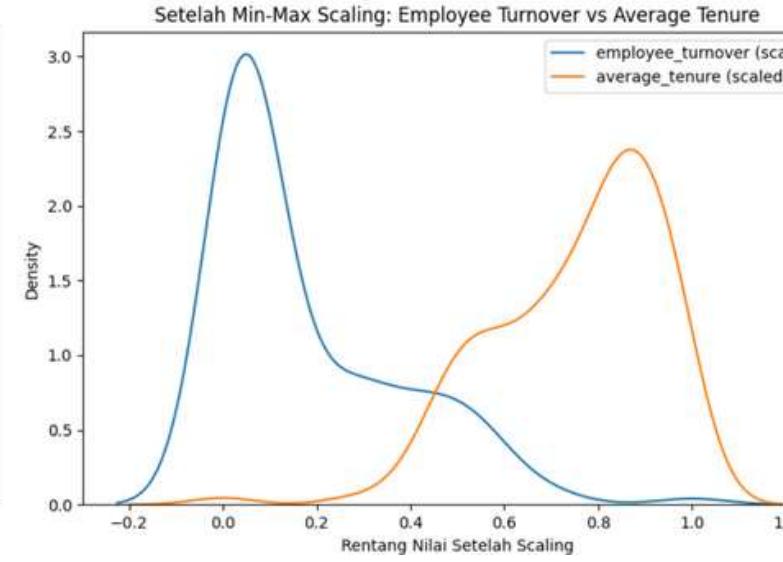
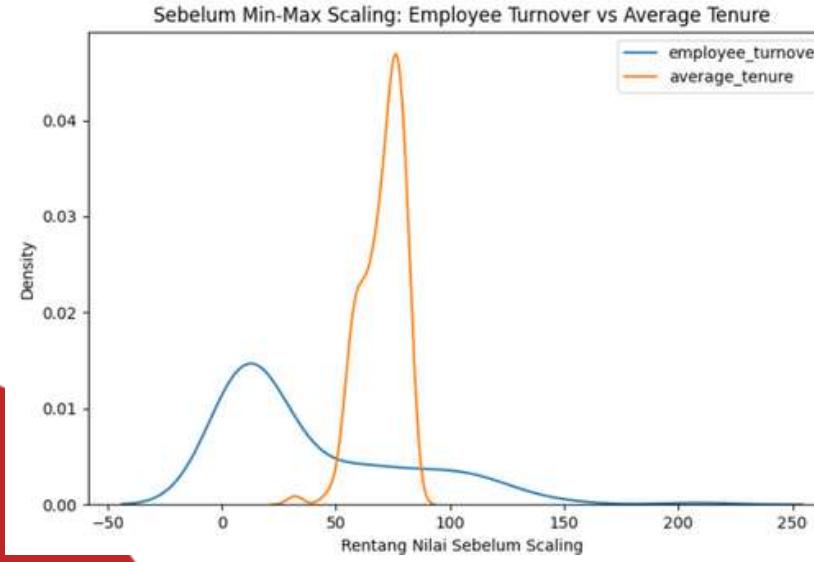
4

Feature Engineering

MinMax Scalling



Revenue_Growth & Growth_potential



Employee_turnover & average_tenure

rd_invesment & Average Salary

4

Feature Engineering

Duplicate Handling



Outlier Handling



Missing Value Handling



Data Encoding



Feature Scalling



5

Supervised Learning

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 167 entries, 0 to 166
Data columns (total 11 columns):
 #   Column           Non-Null Count  Dtype  
--- 
 0   Unnamed: 0        167 non-null    int64  
 1   company          167 non-null    object  
 2   employee_turnover 167 non-null    float64 
 3   revenue_growth   167 non-null    float64 
 4   rd_investment    167 non-null    float64 
 5   operational_cost 167 non-null    float64 
 6   average_salary   167 non-null    int64  
 7   market_volatility 167 non-null    float64 
 8   average_tenure   167 non-null    float64 
 9   growth_potential 167 non-null    float64 
 10  net_profit       167 non-null    int64  
dtypes: float64(7), int64(3), object(1)
memory usage: 14.5+ KB
```

Data Preprocessing

https://colab.research.google.com/drive/1WWuKwqEfKI341KvS7xy_HxnWgHhHC0FF?usp=sharing

Exploratory Data Analyst

Bertujuan untuk mengeksplorasi data atau informasi mengenai datasheet

5

Supervised Learning

Data Preprocessing

Statistik Data Analyst

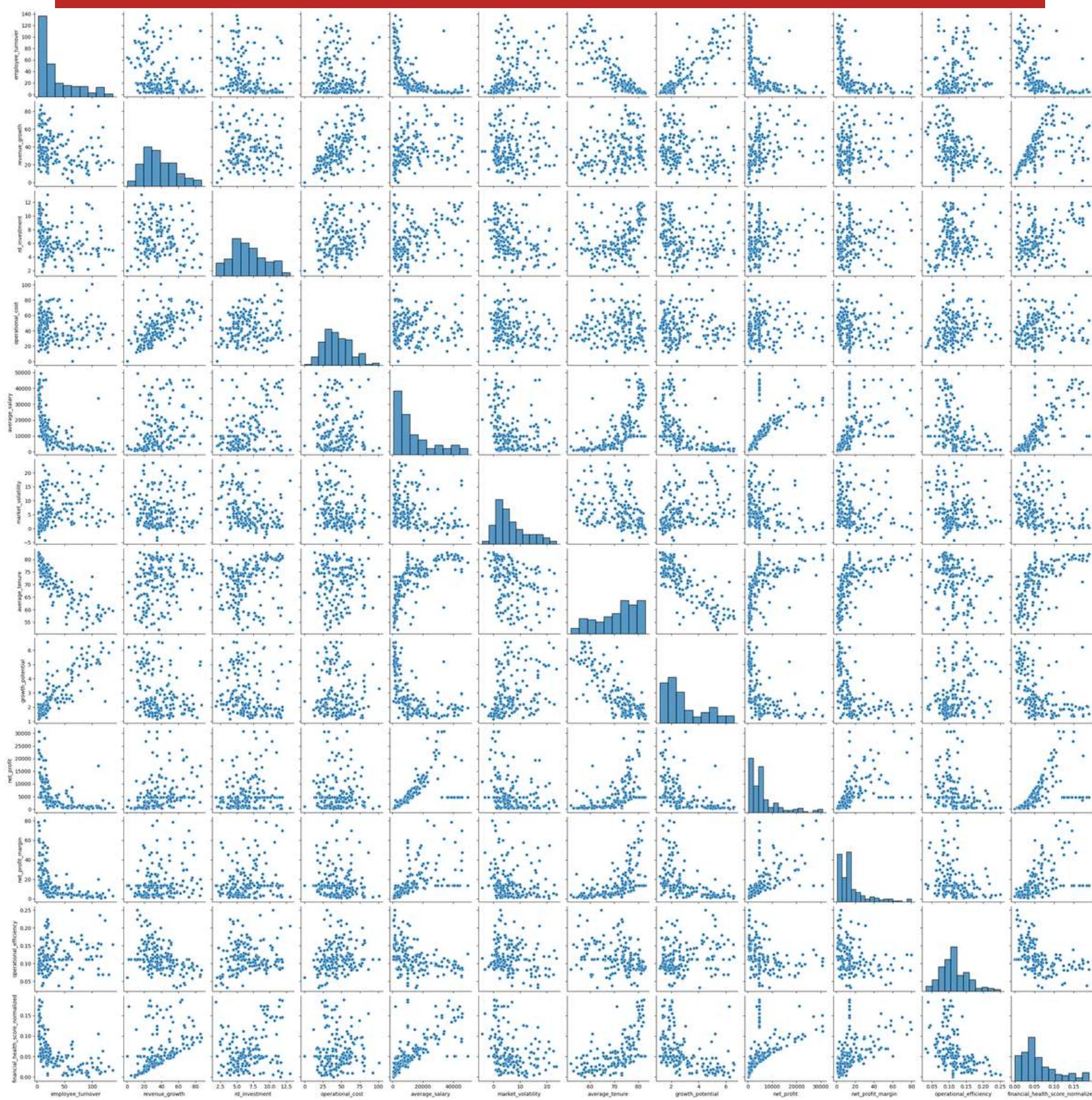
Bertujuan untuk melihat rata-rata, minimum & maksimum semua kolom numerik dari dataset.

	Unnamed: 0	company	employee_turnover	revenue_growth	rd_investment	operational_cost	average_salary	market_volatility	average_tenure	growth_potential	net_profit
count	167.000000	167	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000	167.000000
unique	Nan	167	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan
top	Nan	Company_1	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan
freq	Nan	1	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan	Nan
mean	83.000000	Nan	38.270060	41.108976	6.815689	46.890215	17144.688623	7.781832	70.555689	2.947964	12964.155689
std	48.35287	Nan	40.328931	27.412010	2.746837	24.209589	19278.067698	10.570704	8.893172	1.513848	18328.704809
min	0.000000	Nan	2.600000	0.109000	1.810000	0.065900	609.000000	-4.210000	32.100000	1.150000	231.000000
25%	41.500000	Nan	8.250000	23.800000	4.920000	30.200000	3355.000000	1.810000	65.300000	1.795000	1330.000000
50%	83.000000	Nan	19.300000	35.000000	6.320000	43.300000	9960.000000	5.390000	73.100000	2.410000	4660.000000
75%	124.500000	Nan	62.100000	51.350000	8.600000	58.750000	22800.000000	10.750000	76.800000	3.880000	14050.000000
max	166.000000	Nan	208.000000	200.000000	17.900000	174.000000	125000.000000	104.000000	82.800000	7.490000	105000.000000

5

Supervised Learning

Scatter plot dan histogram



Terdapat beberapa fitur yang memiliki korelasi cukup kuat, diantaranya:

Korelasi Positif:

- average_salary vs revenue_growth → Korelasi positif terlihat jelas.

Artinya, perusahaan dengan rata-rata gaji yang lebih tinggi cenderung memiliki pertumbuhan pendapatan (revenue growth) yang lebih besar. Ini masuk akal karena gaji tinggi bisa mencerminkan kinerja perusahaan yang baik.

Korelasi Negatif:

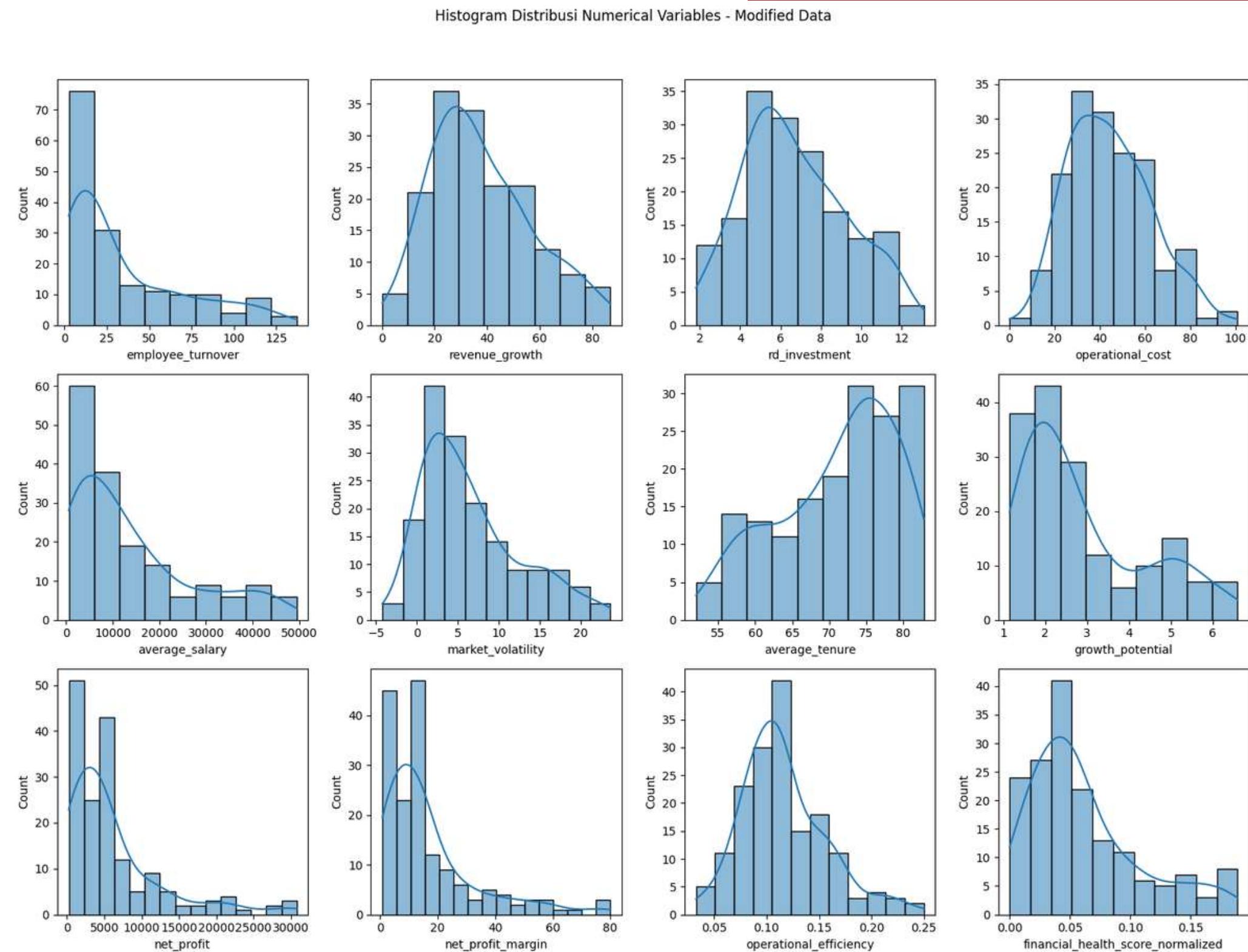
- employees_turnover vs average_salary → Korelasi negatif terlihat cukup konsisten.

Artinya, perusahaan dengan gaji rata-rata tinggi cenderung memiliki turnover karyawan yang lebih rendah. Hal ini sesuai asumsi bahwa karyawan lebih betah jika diberi kompensasi layak.

5

Supervised Learning

Hasil distribusi dari predict variable semua fiture



Korelasi Data

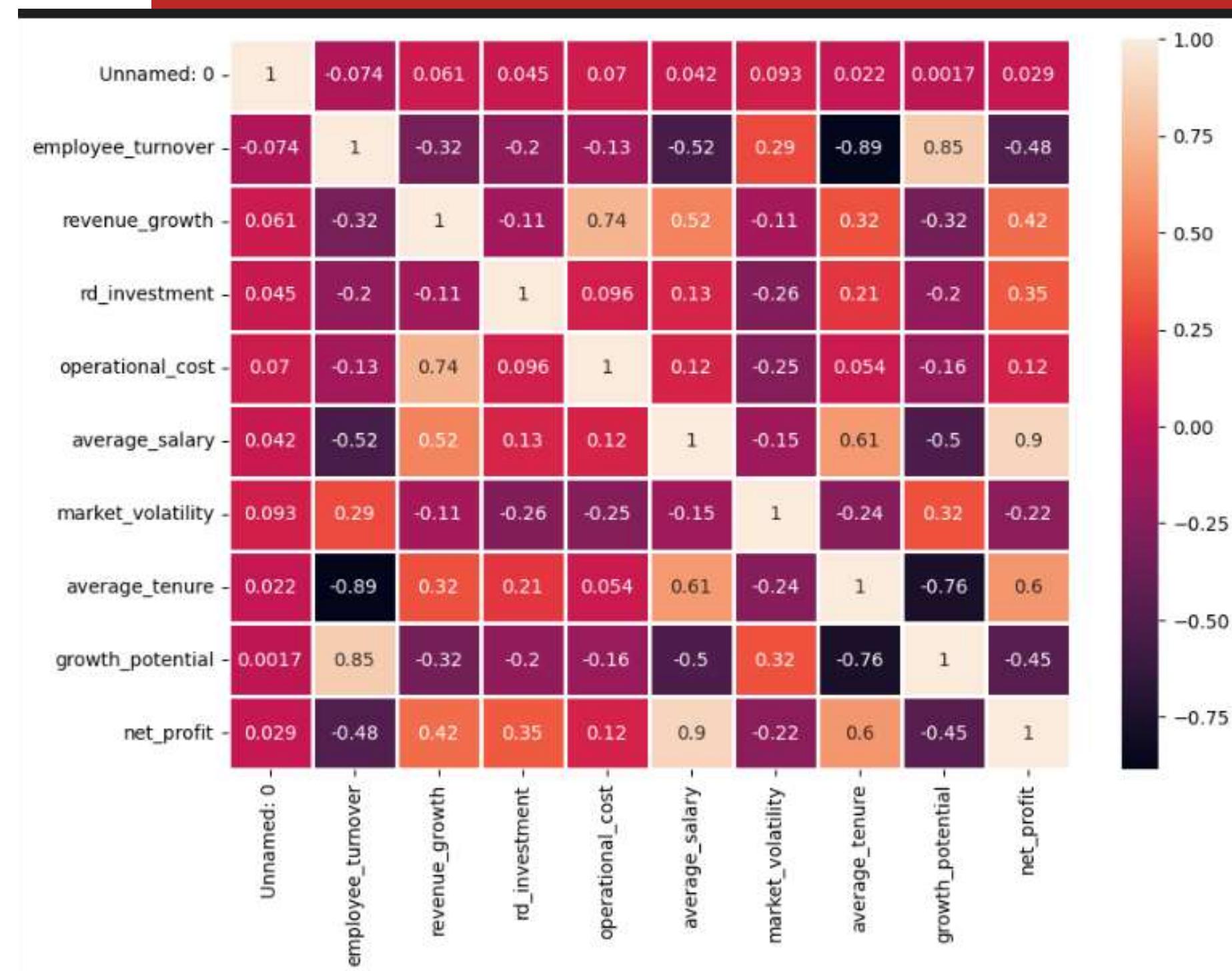
Hasil Korelasi antar fitur:

	Unnamed: 0	employee_turnover	revenue_growth	rd_investment	operational_cost	average_salary	market_volatility	average_tenure	growth_potential	net_profit
Unnamed: 0	1.000000	-0.073652	0.060566	0.044585	0.070232	0.042387	0.093249	0.022213	0.001663	0.029414
employee_turnover	-0.073652	1.000000	-0.318093	-0.200402	-0.127211	-0.524315	0.288276	-0.886676	0.848478	-0.483032
revenue_growth	0.060566	-0.318093	1.000000	-0.114408	0.737381	0.516784	-0.107294	0.316313	-0.320011	0.418725
rd_investment	0.044585	-0.200402	-0.114408	1.000000	0.095717	0.129579	-0.255376	0.210692	-0.196674	0.345966
operational_cost	0.070232	-0.127211	0.737381	0.095717	1.000000	0.122406	-0.246994	0.054391	-0.159048	0.115498
average_salary	0.042387	-0.524315	0.516784	0.129579	0.122406	1.000000	-0.147756	0.611962	-0.501840	0.895571
market_volatility	0.093249	0.288276	-0.107294	-0.255376	-0.246994	-0.147756	1.000000	-0.239705	0.316921	-0.221631
average_tenure	0.022213	-0.886676	0.316313	0.210692	0.054391	0.611962	-0.239705	1.000000	-0.760875	0.600089
growth_potential	0.001663	0.848478	-0.320011	-0.196674	-0.159048	-0.501840	0.316921	-0.760875	1.000000	-0.454910
net_profit	0.029414	-0.483032	0.418725	0.345966	0.115498	0.895571	-0.221631	0.600089	-0.454910	1.000000

- Ada korelasi negatif yang kuat antara employee_turnover dan average_tenure (-0.89), yang masuk akal karena tingkat turnover yang lebih tinggi cenderung berarti masa kerja yang lebih pendek.
- revenue_growth berkorelasi positif dengan operational_cost (0.74), menunjukkan bahwa perusahaan dengan pertumbuhan pendapatan yang lebih tinggi cenderung memiliki biaya operasional yang lebih tinggi.
- average_salary berkorelasi sangat kuat dengan net_profit (0.90), menandakan perusahaan dengan gaji rata-rata yang lebih tinggi cenderung memiliki laba bersih yang lebih besar.
- growth_potential berkorelasi negatif dengan average_tenure (-0.76), menunjukkan bahwa perusahaan dengan potensi pertumbuhan yang lebih tinggi mungkin mengalami turnover karyawan yang lebih tinggi.

meninjau korelasi antar variable menggunakan heatmap.

- drop kolom company
- preparation dataset untuk di training dengan cara memisahkan prediktor dan target variabelnya.



Hasil Evaluasi

5

Hasil Evaluasi dengan melihat MAE, RMSE, R2 Square

1. MAE dan RMSE:

- Nilai MAE dan RMSE yang relatif kecil menunjukkan bahwa secara rata-rata, prediksi model cukup dekat dengan nilai aktual.
- Namun, nilai RMSE yang lebih besar dari MAE mengindikasikan adanya beberapa prediksi yang meleset cukup jauh (karena RMSE lebih sensitif terhadap outlier).

2. R-squared (R²):

- R² square test (0.4) berarti sekitar 40% variasi dalam data target dapat dijelaskan oleh model pada data yang belum pernah dilihat sebelumnya. Ini termasuk kategori moderat, dan bisa menunjukkan bahwa model masih bisa ditingkatkan.
- R² square train (0.744) berarti model mampu menjelaskan 66.1% variasi pada data pelatihan, yang lebih tinggi dibanding data uji.
- Perbedaan antara nilai R² train dan test menunjukkan potensi overfitting ringan, di mana model lebih baik dalam menjelaskan data latih dibandingkan data uji.

Supervised Learning

Model Evaluasi

```
[145] print("MAE:", metrics.mean_absolute_error(y_test,predictions))
      print("RMSE:", np.sqrt(metrics.mean_squared_error(y_test,predictions)))
      print("R-squared test:", round(metrics.r2_score(y_test,predictions),3))
      print("R-squared train:", round(metrics.r2_score(y_train,lm.predict(X_train)),3))
```

→ MAE: 0.023518256040214413
RMSE: 0.03381273797275864
R-squared test: 0.4
R-squared train: 0.744

Thank You