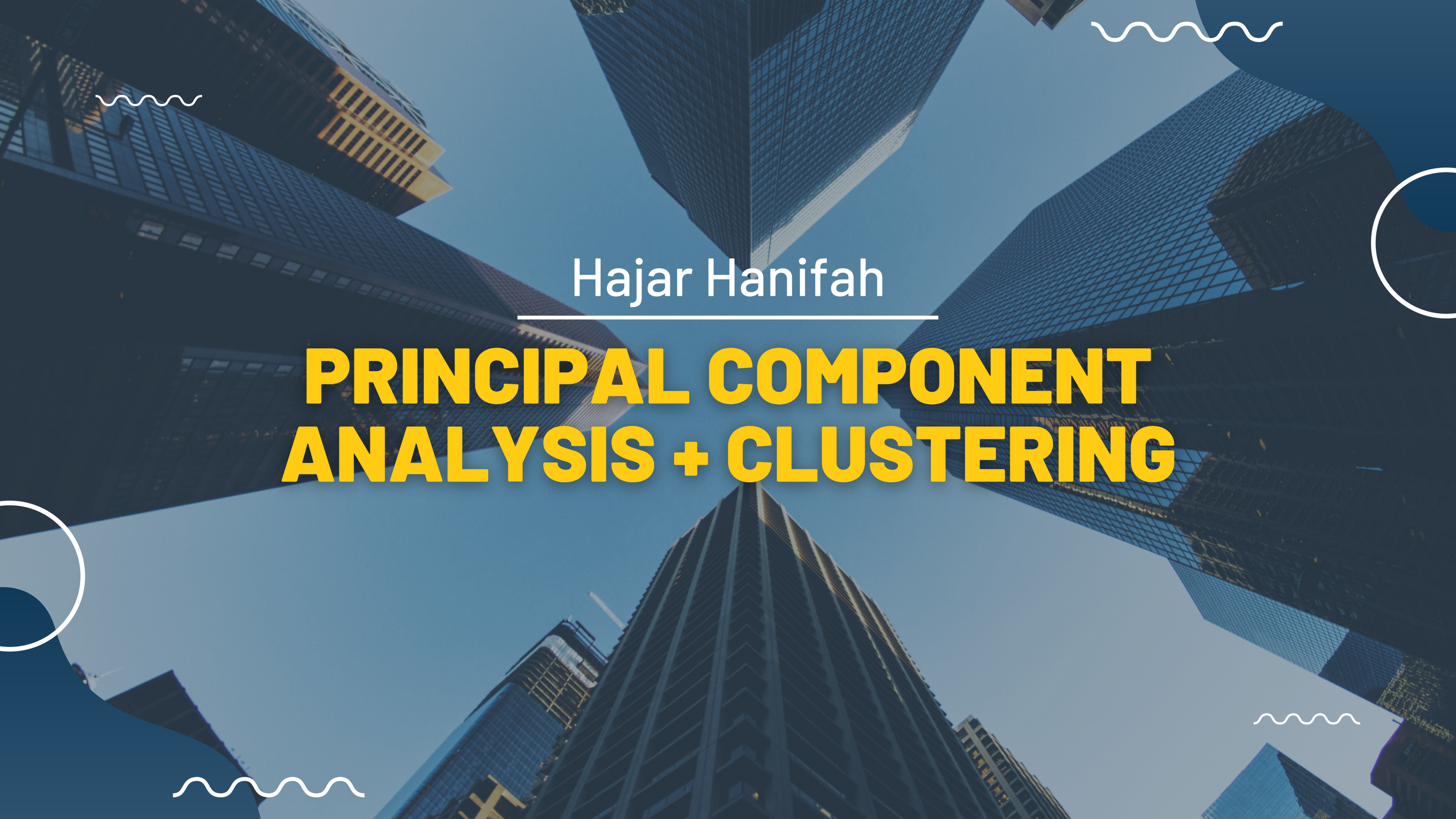


Hajar Hanifah

PRINCIPAL COMPONENT ANALYSIS + CLUSTERING

Business Understanding

Pandemi Covid-19 yang terjadi sejak tahun 2020 memberikan dampak negatif pada multisektor di dunia termasuk Indonesia. Indonesia mengalami pertumbuhan ekonomi negatif pada pada triwulan I dan II di tahun 2020. Sektor Industri merupakan salah satu sektor yang terdampak cukup besar semenjak terjadinya pandemi Covid-19.

PCA

- Teknik yang mentransform high dimension data —> lower dimension data
- dengan mempertahankan data sebanyak banyaknya
    - ex. clustering
        - 20 rasio keuangan
            - bikin 3 rasio yang mewakili 20 rasio
- konsep PCA = Mengurangi dimensi data —> bisa di visualisasikan dengan lebih bagus
- PCA —> digunakan membuat new set dimension
    - dimensi tersebut dilakukan rangking berdasarkan eigen value / eigen vector

Dataset

- Data yang diambil merupakan data statictical financial ratio Q4 2021 untuk emiten di sektor Industri

- Data di dapatkan dari website IDX - Laporan Statistic Financial Data Ratio. https://www.idx.co.id/data-pasar/laporan-statistik/digital- statistic-beta/financial-data-ratio?

Data Understanding

Emiten	Pihak yang melakukan penawaran umum, yaitu penawaran efek yang dilakukan oleh emiten untuk menjual efek kepada masyarakat berdasarkan tata cara yang diatur dalam peraturan undang-undang yang berlaku.
Sales Growth	Kenaikan jumlah penjualan dari tahun ke tahun atau dari waktu ke waktu.
Retturn of Asset (ROA)	Indikator untuk menunjukkan seberapa untuk sebuah perusahaan dibandingkan dengan total asetnya.
Debt to Equity Ratio (DER)	Rasio hutang terhadap ekuitas atau rasio keuangan yang membandingkan jumlah hutang dengan ekuitas

Data Understanding

Category	Hasil Clustering data Emiten sektor Industri - Aneka Industri , Mesin dan Komponen Industri
ROE (Return Of Equity)	Return on equity atau ROE adalah indikator kinerja perusahaan dengan membandingkan laba bersih dan total modal
NPM (Net Profit Margin)	ingkat keuntungan suatu perusahaan dari penjualan atau pendapatan yang diperoleh.

Data Preparation

Siapkan Dataset

- Siapkan dataset berupa data Emiten saham, ROA, DER, dll pada satu file

Cek Baris dan Kolom

- Cek jumlah baris dan kolom, diketahui jumlah baris 29 dan kolom 10

Cek apakah ada data yang bernilai null

- Lakukan pengecekan pada dataset, apakah ada yang bernilai null. Apabila ada data yang bernilai null, isi dengan value 0

Deployment

Bahasa pemrograman yang digunakan adalah Phyton, adapun code dapat di akses di :

https://github.com/Hajarhanifah/big-data/blob/main/multinomial-logistic-regression/multinomial-logistic-regression-2-aneka-industri.ipynb

Data

	No	Industri	Company	Emiten	DER	ROA	Sales Growth	ROE	NPM	Category
0	1	Industrial Machinery & Components	Asahimas Flat Glass Tbk	AMFG	1.320000	0.060000	0.330000	0.130000	0.120000	1
1	2	Industrial Machinery & Components	PT Ateliers Mecaniques D Indonesie Tbk.	AMIN	1.140000	-0.050000	0.410000	-0.120000	-0.130000	1
2	3	Industrial Machinery & Components	PT Arita Prima Indonesia Tbk.	APII	0.490000	0.050000	0.030000	0.070000	0.130000	1
3	4	Industrial Machinery & Components	PT Arkha Jayanti Persada Tbk.	ARKA	3.970000	-0.020000	0.220000	-0.080000	-0.160000	1
4	5	Industrial Machinery & Components	Arwana Citramulia Tbk	ARNA	0.510000	0.210000	0.170000	0.320000	0.250000	4
5	6	Industrial Machinery & Components	Cahayaputra Asa Keramik Tbk	CAKK	0.880000	0.030000	0.300000	0.060000	0.070000	1
6	7	Industrial Machinery & Components	Communication Cable Systems Indonesia Tbk	CCSI	0.310000	0.110000	0.790000	0.140000	0.170000	4
7	8	Industrial Machinery & Components	Citatah Tbk	CTTH	2.280000	-0.050000	-0.170000	-0.160000	-0.540000	1
8	9	Industrial Machinery & Components	Hexindo Adiperkasa Tbk	HEXA	1.380000	0.120000	0.680000	0.290000	0.210000	4
9	10	Industrial Machinery & Components	Sumi Indo Kabel Tbk	IKBI	0.670000	0.010000	0.670000	0.010000	0.010000	1
10	11	Industrial Machinery & Components	Impack Pratama Industri Tbk	IMPC	0.720000	0.080000	0.280000	0.130000	0.130000	1
11	12	Industrial Machinery & Components	Intraco Penta Tbk	INTA	-2.860000	-0.320000	-0.220000	0.000000	-1.990000	3
12	13	Industrial Machinery & Components	Jembo Cable Company Tbk	JECC	1.580000	-0.040000	0.160000	-0.100000	-0.050000	1
13	14	Industrial Machinery & Components	KMI Wire & Cable Tbk	KBLI	0.160000	0.020000	-0.170000	0.020000	0.050000	1
14	15	Industrial Machinery & Components	Kabelindo Murni Tbk	KBLM	0.380000	0.000000	0.360000	-0.010000	-0.010000	1
15	16	Industrial Machinery & Components	Keramika Indonesia Assosiasi Tbk	KIAS	0.180000	-0.010000	0.410000	-0.020000	-0.030000	1
16	17	Industrial Machinery & Components	Kobexindo Tractors Tbk	KOBX	2.790000	0.060000	1.440000	0.240000	0.070000	2
17	18	Industrial Machinery & Components	Kokoh Inti Arebama Tbk	KOIN	8.530000	-0.020000	0.990000	-0.210000	-0.010000	2
18	19	Industrial Machinery & Components	Steadfast Marine Tbk	KPAL	3.150000	-0.020000	0.000000	-0.080000	-0.500000	1
19	20	Industrial Machinery & Components	PT Grand Kartech Tbk	KRAH	16.330000	-0.050000	0.000000	-0.930000	-0.180000	0
20	21	Industrial Machinery & Components	PT Mark Dynamics Indonesia Tbk.	MARK	0.670000	0.310000	1.420000	0.520000	0.400000	4
21	22	Industrial Machinery & Components	Mulia Industrindo Tbk	MLIA	0.910000	0.080000	0.160000	0.150000	0.150000	1
22	23	Industrial Machinery & Components	Supreme Cable Manufacturing & Commerce Tbk	SCCO	0.080000	0.040000	0.160000	0.050000	0.040000	1
23	24	Industrial Machinery & Components	Singaraja Putra Tbk	SINI	4.090000	0.020000	0.290000	0.100000	0.010000	1
24	25	Industrial Machinery & Components	Superkrane Mitra Utama Tbk	SKRN	1.680000	0.010000	-0.280000	0.010000	0.030000	1
25	26	Industrial Machinery & Components	Surya Pertiwi Tbk	SPTO	0.540000	0.060000	0.170000	0.080000	0.110000	1
26	27	Industrial Machinery & Components	Surya Toto Indonesia Tbk	TOTO	0.650000	0.020000	0.140000	0.040000	0.060000	1
27	28	Industrial Machinery & Components	United Tractors Tbk	UNTR	0.590000	0.090000	0.240000	0.150000	0.180000	1
28	29	Industrial Machinery & Components	Voksel Electric Tbk	VOKS	1.910000	-0.060000	-0.150000	-0.180000	-0.140000	1

Data

```
df.describe()
```

	No	DER	ROA	Sales Growth	ROE	NPM	Category
count	29.000000	29.000000	29.000000	29.000000	29.000000	29.000000	29.000000
mean	15.000000	1.897586	0.025517	0.304483	0.021379	-0.053448	1.517241
std	8.514693	3.366741	0.103909	0.431456	0.241568	0.420300	1.121883
min	1.000000	-2.860000	-0.320000	-0.280000	-0.930000	-1.990000	0.000000
25%	8.000000	0.510000	-0.020000	0.030000	-0.080000	-0.050000	1.000000
50%	15.000000	0.880000	0.020000	0.220000	0.040000	0.040000	1.000000
75%	22.000000	1.910000	0.060000	0.410000	0.130000	0.130000	1.000000
max	29.000000	16.330000	0.310000	1.440000	0.520000	0.400000	4.000000

```
df.shape
```

```
(29, 10)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 29 entries, 0 to 28
Data columns (total 10 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   No             29 non-null     int64
 1   Industri       29 non-null     object
 2   Company        29 non-null     object
 3   Emiten         29 non-null     object
 4   DER            29 non-null     float64
 5   ROA            29 non-null     float64
 6   Sales Growth   29 non-null     float64
 7   ROE            29 non-null     float64
 8   NPM            29 non-null     float64
 9   Category       29 non-null     int64
dtypes: float64(5), int64(2), object(3)
memory usage: 2.4+ KB
```

Standarisasi Data

```
from sklearn.preprocessing import StandardScaler
features = ['DER', 'ROA', 'Sales Growth', 'ROE', 'NPM']
# Separating out the features
x = df.loc[:, features].values
# Separating out the target
y = df.loc[:,['Category']].values
# Standardizing the features
data_standard= StandardScaler().fit_transform(x)
```

```
data_standard
```

```
array([[-1.74593085e-01,  3.37729925e-01,  6.01890239e-02,
         4.57607212e-01,  4.19982048e-01],
       [-2.29003587e-01, -7.39628535e-01,  2.48889748e-01,
        -5.95615737e-01, -1.85359870e-01],
       [-4.25485955e-01,  2.39788246e-01, -6.47438690e-01,
         2.04833705e-01,  4.44195725e-01],
       [ 6.26450413e-01, -4.45803500e-01, -1.99274471e-01,
        -4.27100065e-01, -2.58000900e-01],
       [-4.19440343e-01,  1.80685510e+00, -3.17212423e-01,
         1.25805665e+00,  7.34759846e-01],
       [-3.07596534e-01,  4.39048902e-02, -1.05737474e-02,
         1.62704787e-01,  2.98913665e-01],
       [-4.79896456e-01,  8.27438315e-01,  1.14521818e+00,
         4.99736130e-01,  5.41050432e-01],
       [ 1.15596258e-01, -7.39628535e-01, -1.11919050e+00,
        -7.64131409e-01, -1.17812062e+00],
       [-1.56456251e-01,  9.25379993e-01,  8.85754690e-01,
         1.13166990e+00,  6.37905139e-01],
       [-3.71075453e-01, -1.51978466e-01,  8.62167099e-01,
```

- PCA dipengaruhi oleh skala / scale, sehingga diperlukan untuk men skalakan fitur dalam data sebelum menerapkan PCA.
- - gunakan StandardScaler untuk membantu menstandarisasi dataset ke unit skala (mean = 0, variance = 1) untuk mengoptimalkan kinerja algoritma dari Machine Learning
- - Untuk melihat efek negatif dari tidak menskalakan dataset, dapat menggunakan sckit-learn

2.1 PCA Projection dengan 3 Komponen

+ Code + Markdown

- Data original memiliki 5 kolom (untuk X) yaitu DER, ROI, Sales Frowth, ROE, NPM
- Pada bagian ini, proyeksikan data asli yaitu 5 dimensional data menjadi 3 dimensi / komponen

```
]:
pca = PCA(n_components=3)

principal_components = pca.fit_transform(data_standard)

new_X = pd.DataFrame(data = principal_components, columns=['PC1','PC2','PC3'])
```

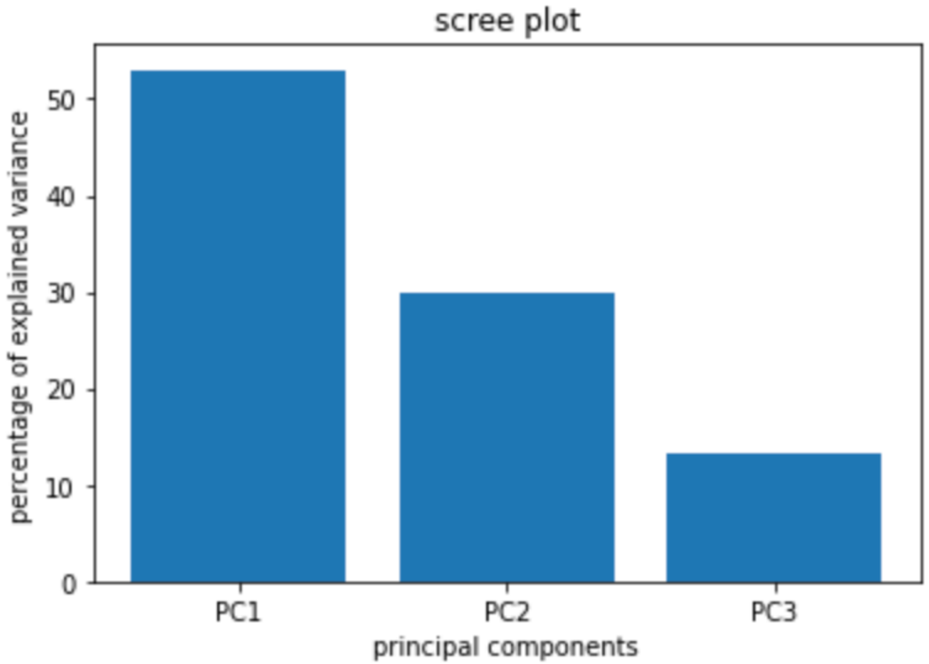
+ Code + Markdown

```
new_X = pd.DataFrame(data = principal_components, columns=['PC1','PC2','PC3'])
new_X.head()
```

	PC1	PC2	PC3
0	-0.676857	-0.098001	-0.210746
1	0.645878	-0.064995	0.414410
2	-0.269784	-0.339275	-0.836648
3	0.800063	0.429666	0.099253
4	-1.943338	-0.311278	-1.018369

data baru yang terbentuk, yaitu data PC1, PC 2 dan PC3

```
plt.bar (x=range(1, len(per_var)+1),height=per_var,tick_label=label)
plt.ylabel('percentage of explained variance')
plt.xlabel('principal components')
plt.title('scree plot')
plt.show()
```

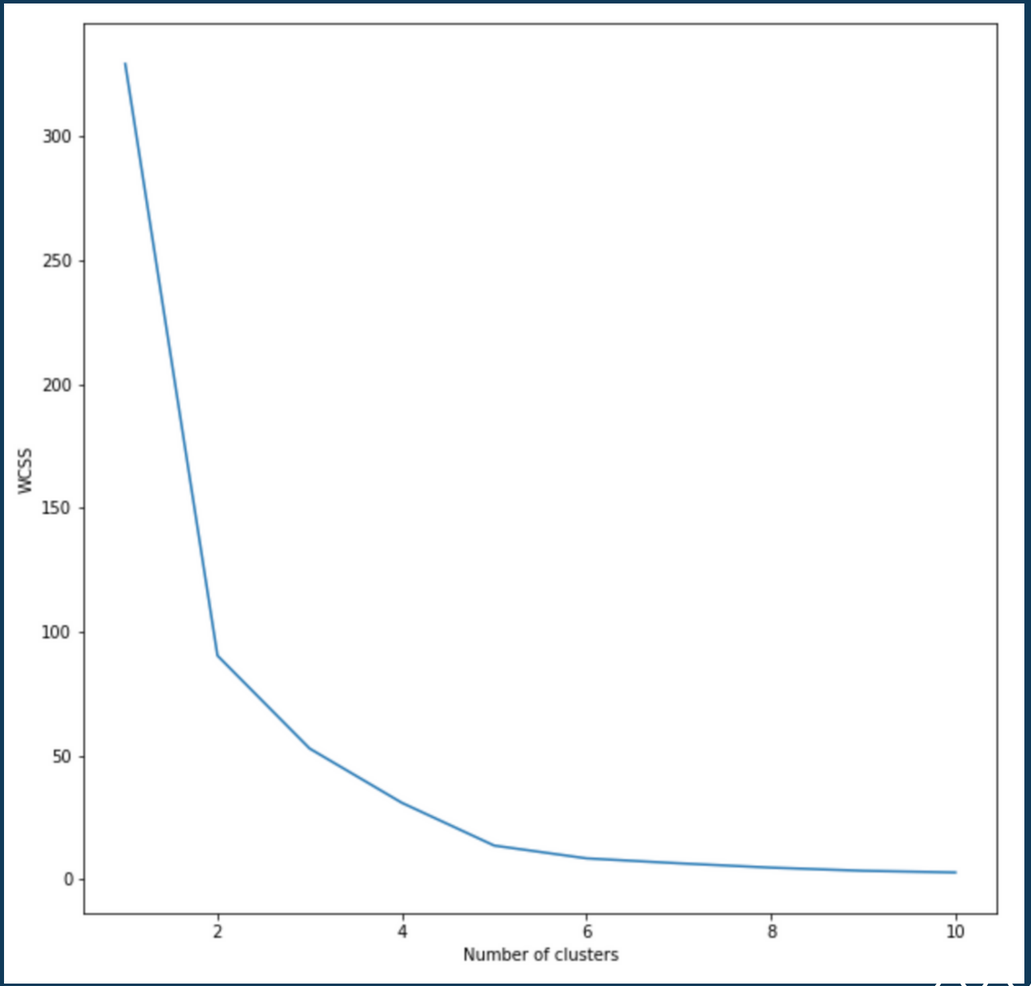


PCA - Clustering

4.1 K-Means Clustering dan PCA

- cari elbow point, untuk menentukan jumlah cluster yang ingin dipertahankan
- pendekatannya dengan cara, ambil titik siku pada grafik

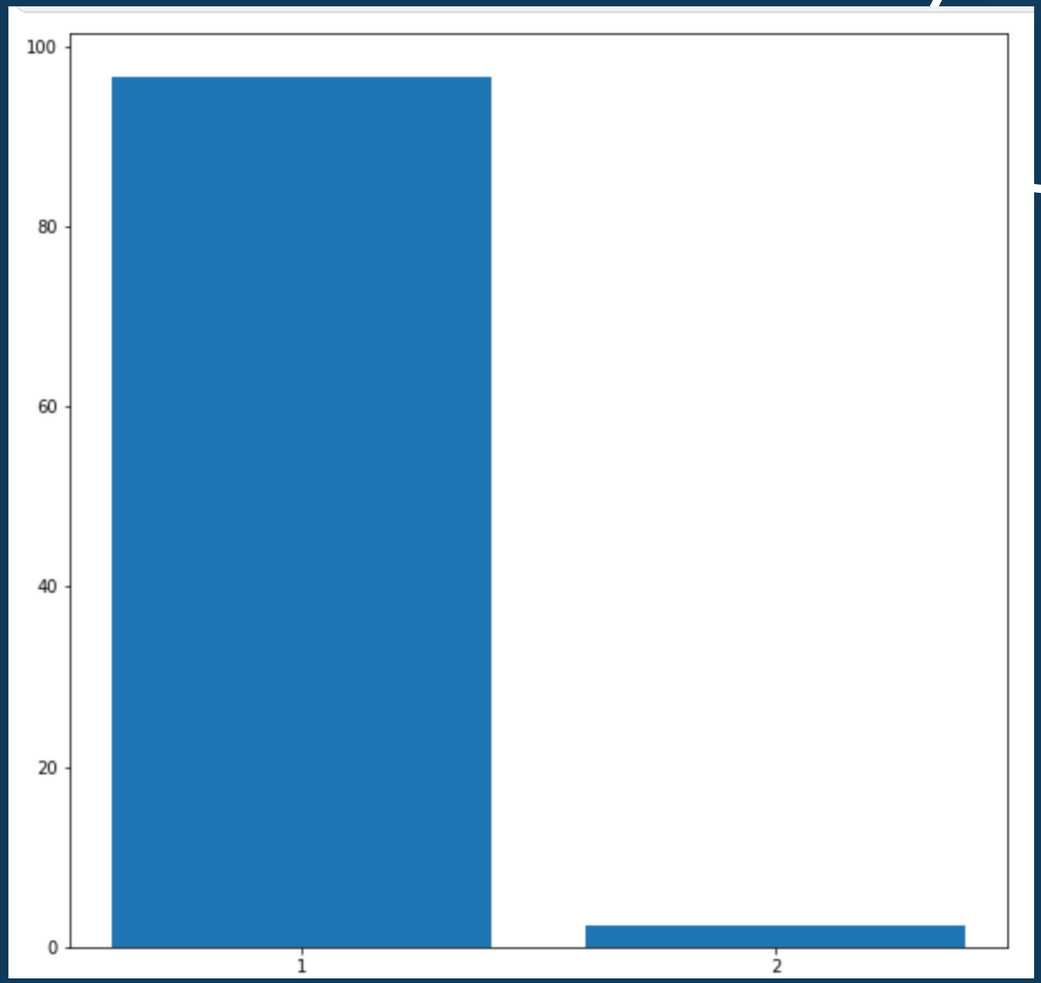
```
wcss = []
for i in range(1,11):
   model = KMeans(n_clusters = i, init = "k-means++")
   model.fit(x)
   wcss.append(model.inertia_)
plt.figure(figsize=(10,10))
plt.plot(range(1,11), wcss)
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```



PCA - Clustering

```
pca = PCA(2)
data = pca.fit_transform(x)
```

```
plt.figure(figsize=(10,10))
var = np.round(pca.explained_variance_ratio_*100, decimals = 1)
lbls = [str(x) for x in range(1,len(var)+1)]
plt.bar(x=range(1,len(var)+1), height = var, tick_label = lbls)
plt.show()
```



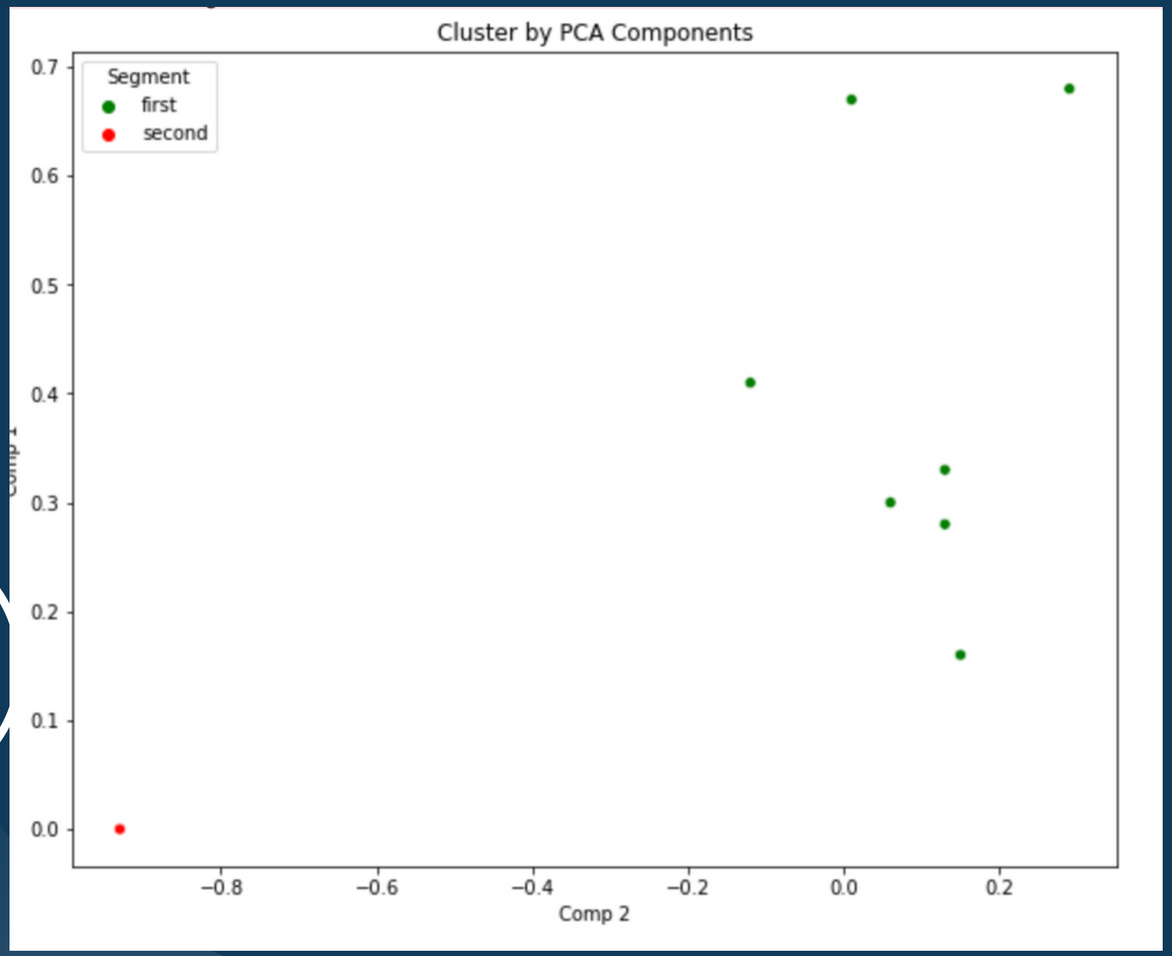
PCA - Clustering

4.2 Analisa Hasil K-Means Clustering dengan PCA

```
#buat dataframe dengan original feature, tambahkan PCA score dan assign clusternya
df_segm_pca_kmeans = pd.concat([df.reset_index(drop = True), pd.DataFrame(x)], axis = 1)
df_segm_pca_kmeans.columns.values[-3:] = ['Comp 1', 'Comp 2', 'Comp 3']
#Tambahkan label pada kolom terakhir tabel
df_segm_pca_kmeans['Segment K-Means PCA'] = model.labels_
df_segm_pca_kmeans
```

	No	Industri	Company	Emiten	DER	ROA	Sales Growth	ROE	NPM	Category	0	1	Comp 1	Comp 2	Comp 3	Segment K-Means PCA
0	1	Industrial Machinery & Components	Asahimas Flat Glass Tbk	AMFG	1.32	0.06	0.33	0.13	0.12	1	1.32	0.06	0.33	0.13	0.12	0
1	2	Industrial Machinery & Components	PT Ateliers Mecaniques D Indonesie Tbk.	AMIN	1.14	-0.05	0.41	-0.12	-0.13	1	1.14	-0.05	0.41	-0.12	-0.13	0
2	3	Industrial Machinery & Components	PT Arita Prima Indonesia Tbk.	APII	0.49	0.05	0.03	0.07	0.13	1	0.49	0.05	0.03	0.07	0.13	9
3	4	Industrial Machinery & Components	PT Arkha Jayanti Persada Tbk.	ARKA	3.97	-0.02	0.22	-0.08	-0.16	1	3.97	-0.02	0.22	-0.08	-0.16	2
4	5	Industrial Machinery & Components	Arwana Citramulia Tbk	ARNA	0.51	0.21	0.17	0.32	0.25	4	0.51	0.21	0.17	0.32	0.25	9
5	6	Industrial Machinery & Components	Cahayaputra Asa Keramik Tbk	CAKK	0.88	0.03	0.30	0.06	0.07	1	0.88	0.03	0.30	0.06	0.07	0
6	7	Industrial Machinery & Components	Communication Cable Systems Indonesia Tbk	CCSI	0.31	0.11	0.79	0.14	0.17	4	0.31	0.11	0.79	0.14	0.17	7
7	8	Industrial Machinery & Components	Citatah Tbk	CTTH	2.28	-0.05	-0.17	-0.16	-0.54	1	2.28	-0.05	-0.17	-0.16	-0.54	5
8	9	Industrial Machinery & Components	Hexindo Adiperkasa Tbk	HEXA	1.38	0.12	0.68	0.29	0.21	4	1.38	0.12	0.68	0.29	0.21	0
9	10	Industrial Machinery & Components	Sumi Indo Kabel Tbk	IKBI	0.67	0.01	0.67	0.01	0.01	1	0.67	0.01	0.67	0.01	0.01	0
10	11	Industrial Machinery & Components	Impack Pratama Industri Tbk	IMPC	0.72	0.08	0.28	0.13	0.13	1	0.72	0.08	0.28	0.13	0.13	0
11	12	Industrial Machinery & Components	Intraco Penta Tbk	INTA	-2.86	-0.32	-0.22	0.00	-1.99	3	-2.86	-0.32	-0.22	0.00	-1.99	4
12	13	Industrial Machinery & Components	Jembo Cable Company Tbk	JECC	1.58	-0.04	0.16	-0.10	-0.05	1	1.58	-0.04	0.16	-0.10	-0.05	5
13	14	Industrial Machinery & Components	KMI Wire & Cable Tbk	KBLI	0.16	0.02	-0.17	0.02	0.05	1	0.16	0.02	-0.17	0.02	0.05	9
14	15	Industrial Machinery & Components	Kabelindo Murni Tbk	KBLM	0.38	0.00	0.36	-0.01	-0.01	1	0.38	0.00	0.36	-0.01	-0.01	9
15	16	Industrial Machinery & Components	Keramika Indonesia Assosiasi Tbk	KIAS	0.18	-0.01	0.41	-0.02	-0.03	1	0.18	-0.01	0.41	-0.02	-0.03	9

Visualisasi



PCA = untuk menentukan komponen yang paling penting.
Dengan cara ini, kita dapat benar-benar yakin bahwa dua komponen pertama menjelaskan lebih banyak varians.

Kesimpulan

It can be helpful to send out a digital copy to potential customers or possible investors, along with a thank you

Terima Kasih

email : hajar.hanifah@gmail.com
github : github.com/hajarhanifah
kaggle : hajarhanifah

Visit our website for further inquiries
www.reallygreatsite.com


LET'S WORK TOGETHER
Contact Info:
123-456-7890
hello@reallygreatsite.com
123 Anywhere St., Any City