Load Data

```
import pandas as pd
import numpy as np
import seaborn as sns

df = pd.read_csv ('Data.csv', delimiter=';')

df
```

```
    Provinsi       Kab/Kota  \
0       ACEH       Simeulue
1       ACEH   Aceh Singkil
2       ACEH   Aceh Selatan
3       ACEH  Aceh Tenggara
4       ACEH     Aceh Timur
..       ...             ...
994      NaN             NaN
995      NaN             NaN
996      NaN             NaN
997      NaN             NaN
998      NaN             NaN

    Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen)  \
0                                                   18,98
1                                                   20,36
2                                                   13,18
3                                                   13,41
4                                                   14,45
..                                                    ...
994                                                   NaN
995                                                   NaN
996                                                   NaN
997                                                   NaN
998                                                   NaN

    Rata-rata Lama Sekolah Penduduk 15+ (Tahun)  \
0                                          9,48
1                                          8,68
2                                          8,88
3                                          9,67
4                                          8,21
..                                          ...
994                                        NaN
995                                        NaN
996                                        NaN
997                                        NaN
998                                        NaN

     Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)  \
```

```
0                                                  7148.0
1                                                  8776.0
2                                                  8180.0
3                                                  8030.0
4                                                  8577.0
..                                                    ...
994                                                   NaN
995                                                   NaN
996                                                   NaN
997                                                   NaN
998                                                   NaN

    Indeks Pembangunan Manusia Umur Harapan Hidup (Tahun)  \
0                         66,41                         65,28
1                         69,22                         67,43
2                         67,44                          64,4
3                         69,44                         68,22
4                         67,83                         68,74
..                          ...                           ...
994                         NaN                           NaN
995                         NaN                           NaN
996                         NaN                           NaN
997                         NaN                           NaN
998                         NaN                           NaN

    Persentase rumah tangga yang memiliki akses terhadap sanitasi
layak  \
0                                                  71,56

1                                                  69,56

2                                                  62,55

3                                                  62,71

4                                                  66,75

..                                                    ...

994                                                   NaN

995                                                   NaN

996                                                   NaN

997                                                   NaN

998                                                   NaN


    Persentase rumah tangga yang memiliki akses terhadap air minum
```

```
layak  \
0                                                  87,45
1                                                  78,58
2                                                  79,65
3                                                  86,71
4                                                  83,16
..                                                   ...
994                                                  NaN
995                                                  NaN
996                                                  NaN
997                                                  NaN
998                                                  NaN


    Tingkat Pengangguran Terbuka Tingkat Partisipasi Angkatan Kerja  \
0                           5,71                                71,15
1                           8,36                                62,85
2                           6,46                                60,85
3                           6,43                                69,62
4                           7,13                                59,48
..                           ...                                   ...
994                          NaN                                  NaN
995                          NaN                                  NaN
996                          NaN                                  NaN
997                          NaN                                  NaN
998                          NaN                                  NaN

     PDRB atas Dasar Harga Konstan menurut Pengeluaran (Rupiah)  \
0                                                 1648096.0
1                                                 1780419.0
2                                                 4345784.0
3                                                 3487157.0
4                                                 8433526.0
..                                                      ...
994                                                   NaN
995                                                   NaN
996                                                   NaN
997                                                   NaN
998                                                   NaN

     Klasifikasi Kemiskinan
```

```
0                        0.0
1                        1.0
2                        0.0
3                        0.0
4                        0.0
..                       ...
994                      NaN
995                      NaN
996                      NaN
997                      NaN
998                      NaN

[999 rows x 13 columns]
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 999 entries, 0 to 998
Data columns (total 13 columns):
 #   Column
Non-Null Count  Dtype
---  ------
--------------  -----
 0   Provinsi
514 non-null    object
 1   Kab/Kota
514 non-null    object
 2   Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen)
514 non-null    object
 3   Rata-rata Lama Sekolah Penduduk 15+ (Tahun)
514 non-null    object
 4   Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)
514 non-null    float64
 5   Indeks Pembangunan Manusia
514 non-null    object
 6   Umur Harapan Hidup (Tahun)
514 non-null    object
 7   Persentase rumah tangga yang memiliki akses terhadap sanitasi
layak   514 non-null    object
 8   Persentase rumah tangga yang memiliki akses terhadap air minum
layak  514 non-null    object
 9   Tingkat Pengangguran Terbuka
514 non-null    object
 10  Tingkat Partisipasi Angkatan Kerja
514 non-null    object
 11  PDRB atas Dasar Harga Konstan menurut Pengeluaran (Rupiah)
514 non-null    float64
 12  Klasifikasi Kemiskinan
514 non-null    float64
```

```
dtypes: float64(3), object(10)
memory usage: 101.6+ KB
```

```
df.isnull().sum()
```

```
Provinsi
485
Kab/Kota
485
Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen)
485
Rata-rata Lama Sekolah Penduduk 15+ (Tahun)
485
Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)
485
Indeks Pembangunan Manusia
485
Umur Harapan Hidup (Tahun)
485
Persentase rumah tangga yang memiliki akses terhadap sanitasi layak
485
Persentase rumah tangga yang memiliki akses terhadap air minum layak
485
Tingkat Pengangguran Terbuka
485
Tingkat Partisipasi Angkatan Kerja
485
PDRB atas Dasar Harga Konstan menurut Pengeluaran (Rupiah)
485
Klasifikasi Kemiskinan
485
dtype: int64
```

Pre - Processing data Mengubah tipe data objek menjadi float

```
df['Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen)']
= df['Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota
(Persen)'].str.replace(',', '.').astype(float)

df['Rata-rata Lama Sekolah Penduduk 15+ (Tahun)'] = df['Rata-rata Lama
Sekolah Penduduk 15+ (Tahun)'].str.replace(',', '.').astype(float)

df['Indeks Pembangunan Manusia'] = df['Indeks Pembangunan
Manusia'].str.replace(',', '.').astype(float)

df['Umur Harapan Hidup (Tahun)'] = df['Umur Harapan Hidup
(Tahun)'].str.replace(',', '.').astype(float)
```

```
df['Persentase rumah tangga yang memiliki akses terhadap sanitasi layak'] = df['Persentase rumah tangga yang memiliki akses terhadap sanitasi layak'].str.replace(',', '.').astype(float)

df['Persentase rumah tangga yang memiliki akses terhadap air minum layak'] = df['Persentase rumah tangga yang memiliki akses terhadap air minum layak'].str.replace(',', '.').astype(float)

df['Tingkat Pengangguran Terbuka'] = df['Tingkat Pengangguran Terbuka'].str.replace(',', '.').astype(float)

df['Tingkat Partisipasi Angkatan Kerja'] = df['Tingkat Partisipasi Angkatan Kerja'].str.replace(',', '.').astype(float)
```

Menghapus missing value

```
df = df.dropna(how='all')
```

Mengubah tipe data float menjadi integer

```
df['Klasifikasi Kemiskinan'] = df['Klasifikasi Kemiskinan'].astype(int)
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_11496\3882940112.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df['Klasifikasi Kemiskinan'] = df['Klasifikasi Kemiskinan'].astype(int)
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 514 entries, 0 to 513
Data columns (total 13 columns):
 #   Column                                                  Non-Null Count  Dtype
---  ------                                                  --------------  -----
 0   Provinsi                                                514 non-null    object
 1   Kab/Kota                                                514 non-null    object
 2   Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen)  514 non-null    float64
 3   Rata-rata Lama Sekolah Penduduk 15+ (Tahun)
```

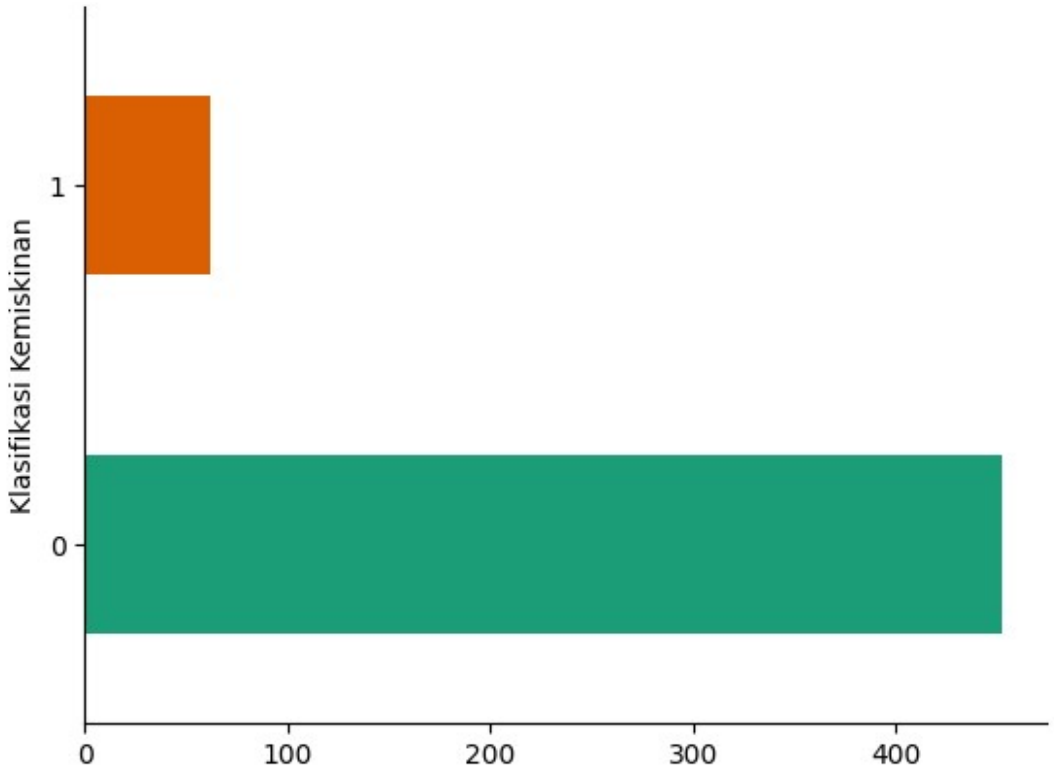
```
514 non-null    float64
 4   Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)
514 non-null    float64
 5   Indeks Pembangunan Manusia
514 non-null    float64
 6   Umur Harapan Hidup (Tahun)
514 non-null    float64
 7   Persentase rumah tangga yang memiliki akses terhadap sanitasi
layak   514 non-null    float64
 8   Persentase rumah tangga yang memiliki akses terhadap air minum
layak  514 non-null    float64
 9   Tingkat Pengangguran Terbuka
514 non-null    float64
 10  Tingkat Partisipasi Angkatan Kerja
514 non-null    float64
 11  PDRB atas Dasar Harga Konstan menurut Pengeluaran (Rupiah)
514 non-null    float64
 12  Klasifikasi Kemiskinan
514 non-null    int32
dtypes: float64(10), int32(1), object(2)
memory usage: 54.2+ KB
```

Explorary data

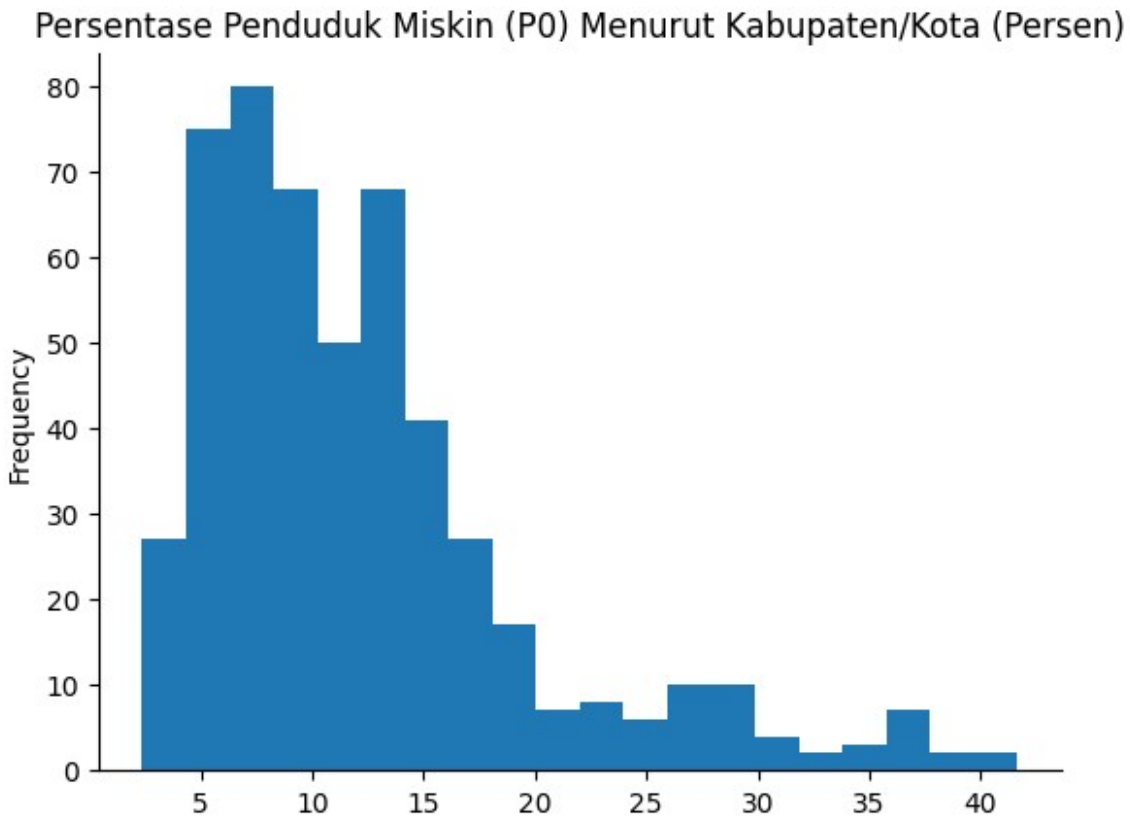
Distribusi

```
from matplotlib import pyplot as plt
import seaborn as sns

df.groupby('Klasifikasi Kemiskinan').size().plot(kind='barh',
color=sns.palettes.mpl_palette('Dark2'))
plt.gca().spines[['top', 'right',]].set_visible(False)
```



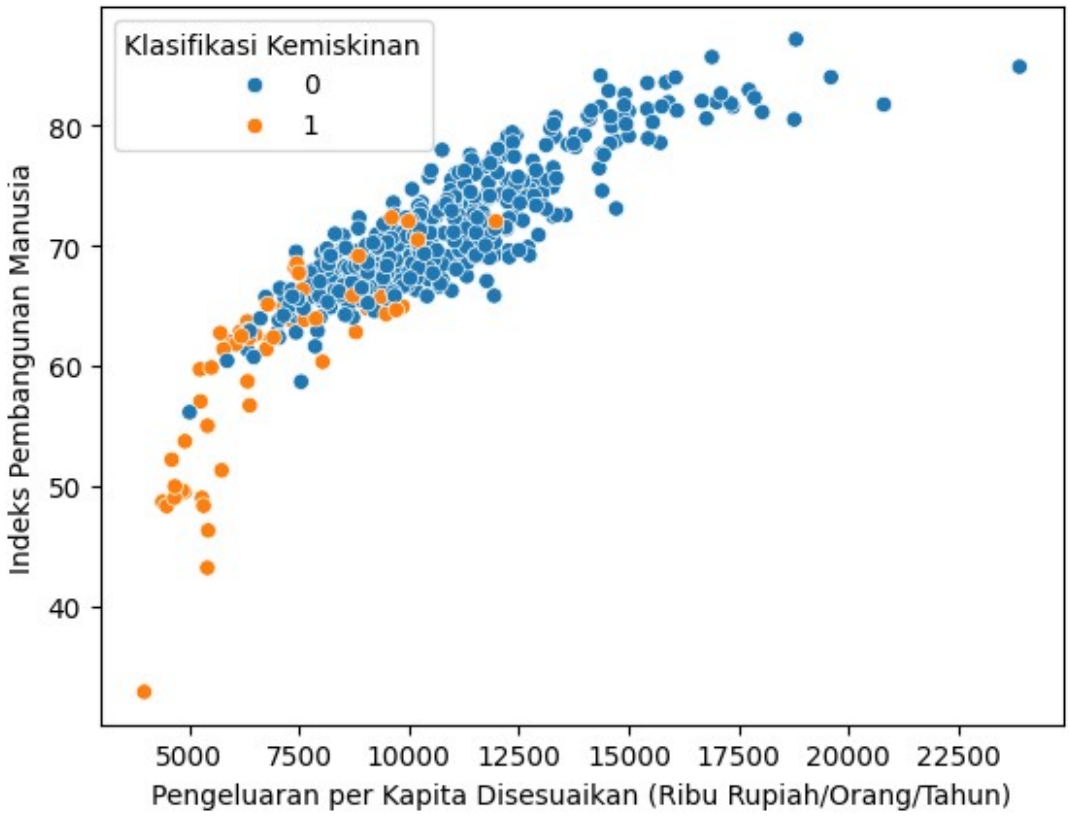

```
df['Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota
(Persen)'].plot(kind='hist', bins=20, title='Persentase Penduduk
Miskin (P0) Menurut Kabupaten/Kota (Persen)')
plt.gca().spines[['top', 'right',]].set_visible(False)
```



Hubungan Variabel

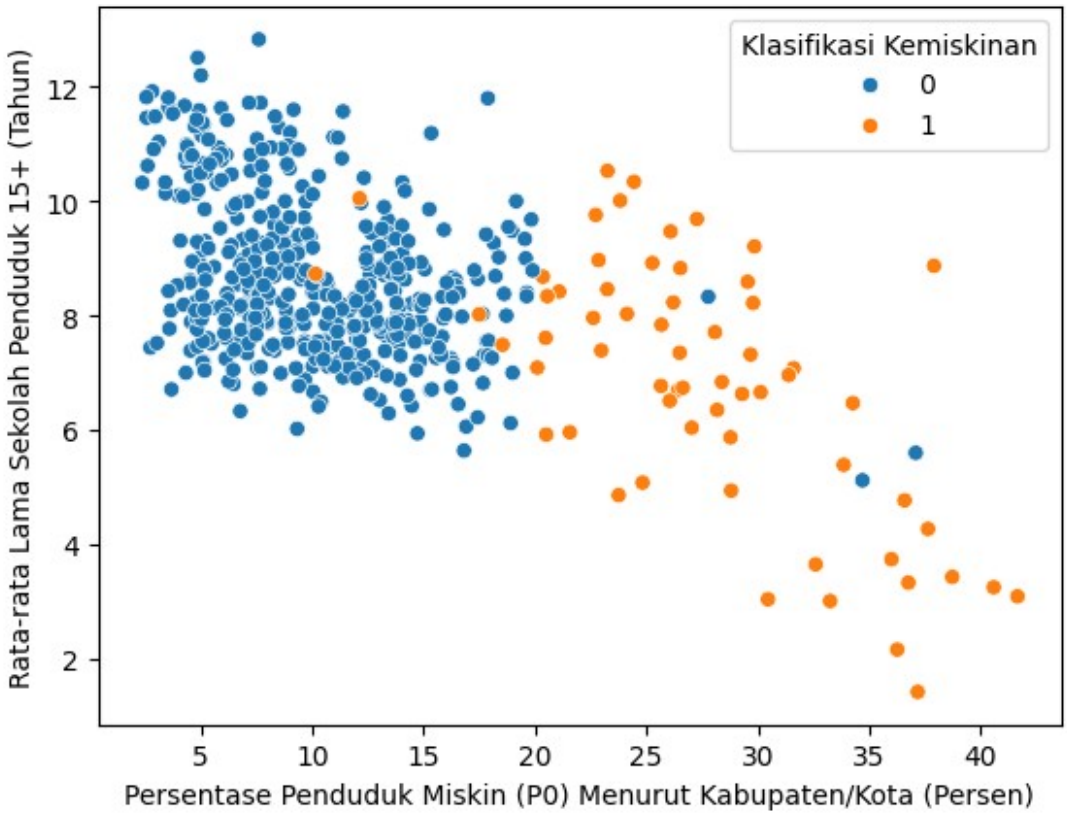
```
sns.scatterplot(x=df['Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)'], y=df['Indeks Pembangunan Manusia'], hue=df['Klasifikasi Kemiskinan'])
```

```
<Axes: xlabel='Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)', ylabel='Indeks Pembangunan Manusia'>
```



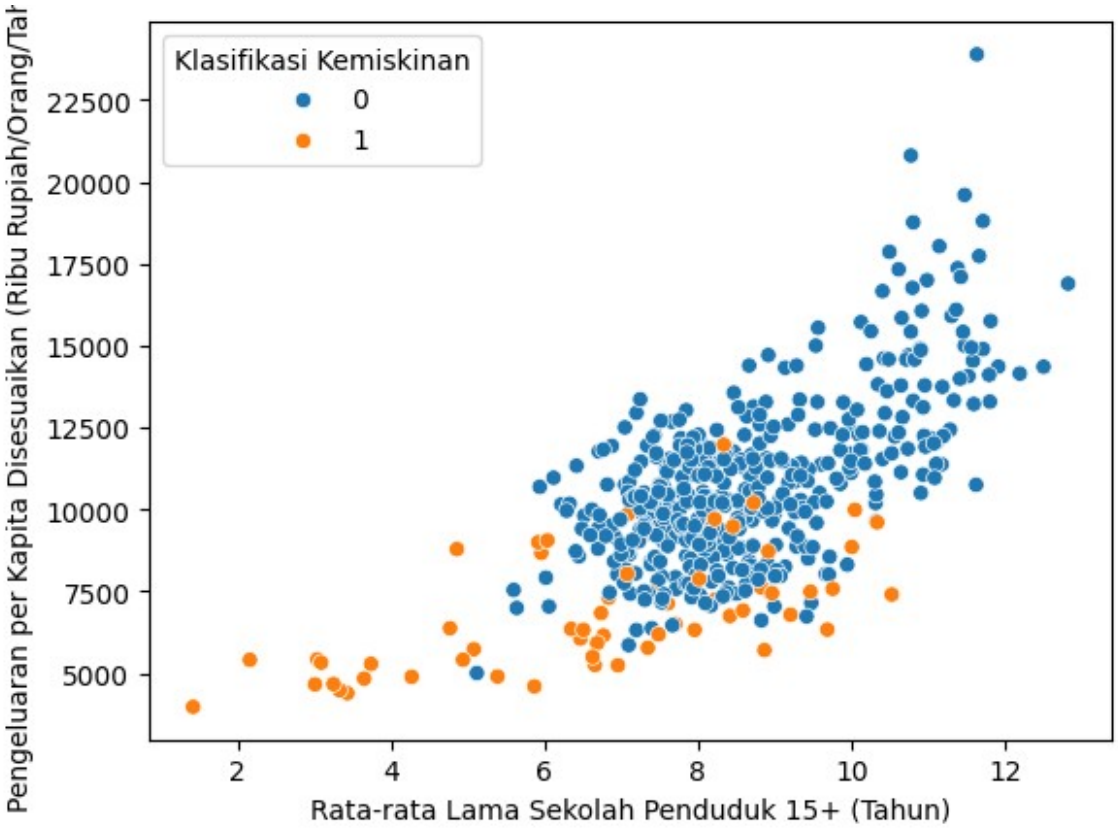

```
sns.scatterplot(x=df['Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen)'], y=df['Rata-rata Lama Sekolah Penduduk 15+ (Tahun)'], hue=df['Klasifikasi Kemiskinan'])
```

```
<Axes: xlabel='Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen)', ylabel='Rata-rata Lama Sekolah Penduduk 15+ (Tahun)'>
```



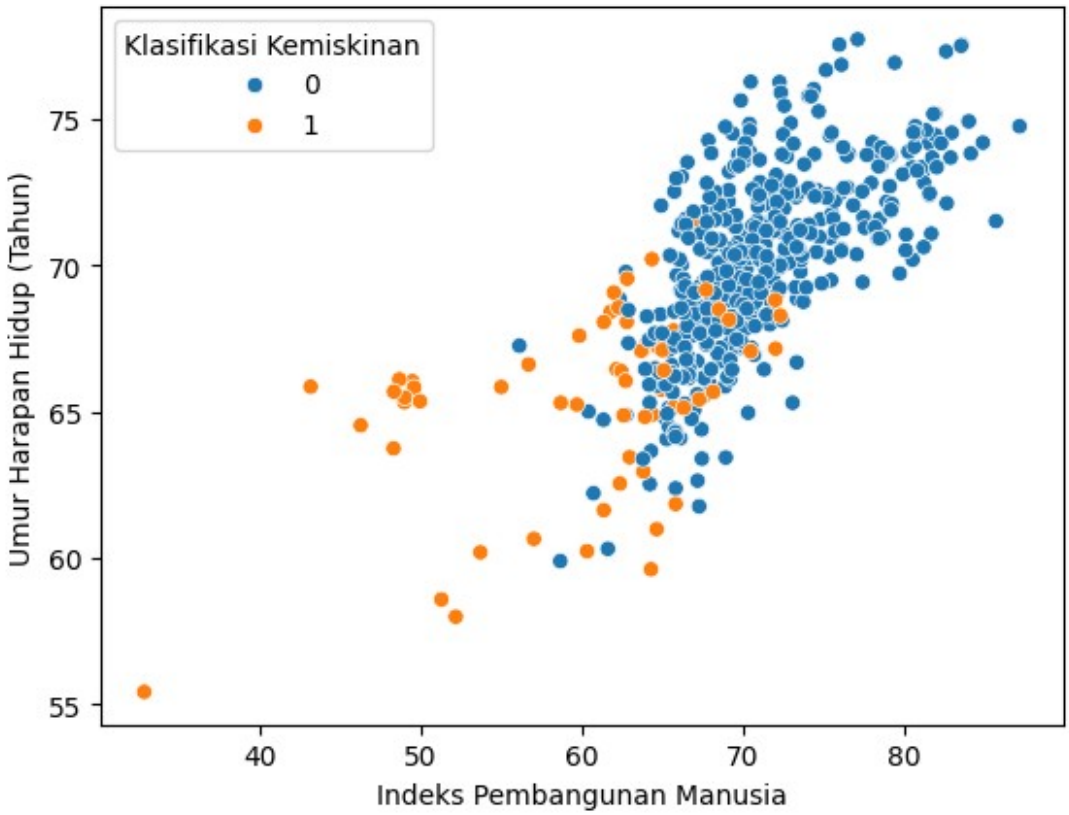

```
sns.scatterplot(x=df['Rata-rata Lama Sekolah Penduduk 15+ (Tahun)'],
y=df['Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)'],
hue=df['Klasifikasi Kemiskinan'])
```

```
<Axes: xlabel='Rata-rata Lama Sekolah Penduduk 15+ (Tahun)',
ylabel='Pengeluaran per Kapita Disesuaikan (Ribu Rupiah/Orang/Tahun)'>
```




```
sns.scatterplot(x=df['Indeks Pembangunan Manusia'], y=df['Umur Harapan Hidup (Tahun)'], hue=df['Klasifikasi Kemiskinan'])
```

```
<Axes: xlabel='Indeks Pembangunan Manusia', ylabel='Umur Harapan Hidup (Tahun)'>
```



Insight satu hal yang menarik pada data ini ketika melakukan exploratory data dengan visualisasi adalah hubungan fitur antara Persentase Penduduk Miskin (P0) Menurut Kabupaten/Kota (Persen) dengan Rata-rata Lama Sekolah Penduduk 15+ (Tahun). pada hasil visualisasi dua fitur ini hampir dapat diklasifikasikan dengan cara klustering, maka dari hasil tersebut mari mencoba melakukan klasifikasi dengan K-MEANS.

K-Means

```
from sklearn.cluster import KMeans

X_cluster = df[['Persentase Penduduk Miskin (P0) Menurut
Kabupaten/Kota (Persen)', 'Rata-rata Lama Sekolah Penduduk 15+
(Tahun)']]

kmeans = KMeans(n_clusters=2, random_state=21)
df['Klaster K-Means'] = kmeans.fit_predict(X_cluster)
```

```
C:\Users\ASUS\AppData\Local\Temp\ipykernel_11496\3863722644.py:2:
SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation:
https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#
```

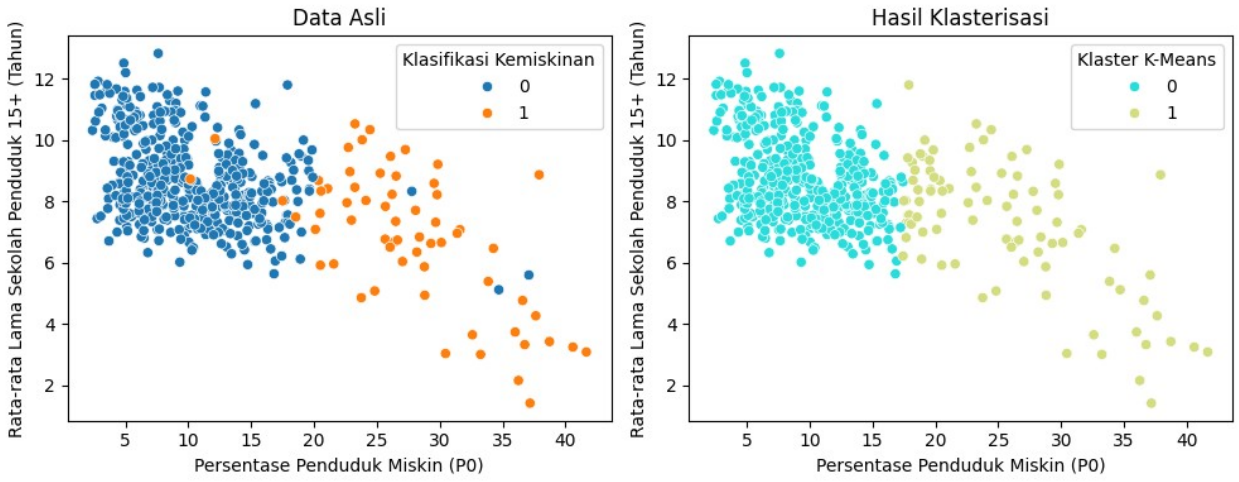
```
returning-a-view-versus-a-copy
  df['Klaster K-Means'] = kmeans.fit_predict(X_cluster)
```

Evaluasi Hasil K-means klustering

```
# Scatterplot data asli
plt.figure(figsize=(10, 4))
plt.subplot(1, 2, 1)  # 1 baris, 2 kolom, plot pertama
sns.scatterplot(x=df['Persentase Penduduk Miskin (P0) Menurut
Kabupaten/Kota (Persen)'],
                y=df['Rata-rata Lama Sekolah Penduduk 15+ (Tahun)'],
                hue=df['Klasifikasi Kemiskinan'])
plt.title('Data Asli')
plt.xlabel('Persentase Penduduk Miskin (P0)')
plt.ylabel('Rata-rata Lama Sekolah Penduduk 15+ (Tahun)')

# Scatterplot hasil klasterisasi
plt.subplot(1, 2, 2)  # 1 baris, 2 kolom, plot kedua
sns.scatterplot(x=df['Persentase Penduduk Miskin (P0) Menurut
Kabupaten/Kota (Persen)'],
                y=df['Rata-rata Lama Sekolah Penduduk 15+ (Tahun)'],
                hue=df['Klaster K-Means'], palette='rainbow')
plt.title('Hasil Klasterisasi')
plt.xlabel('Persentase Penduduk Miskin (P0)')
plt.ylabel('Rata-rata Lama Sekolah Penduduk 15+ (Tahun)')

plt.tight_layout()  # Untuk memastikan layout plot rapi
plt.show()
```




```
from sklearn.metrics import accuracy_score
# Menghitung akurasi K-Means
accuracy_nb = accuracy_score(df['Klasifikasi Kemiskinan'], df['Klaster
```

```
K-Means'])
print(f"Akurasi K-Means: {accuracy_nb}")
```

```
Akurasi K-Means: 0.9396887159533074
```