Kode program yang akan diajarkan di sini bisa diunduh di tautan [berikut ini](https://colab.research.google.com/drive/1WeU2lfVQDZgt8_ldyzbBJDH-VfYwZEqf?usp=sharing). Silahkan buat salinan dari kode ini dan jalankan di Google Colab Anda.

**Tujuan**

Pada latihan kali ini, Anda akan belajar membuat model *unsupervised learning*dengan teknik *K-Means Clustering*.

**Tahapan Latihan**

Tahapan pada latihan kali ini adalah sebagai berikut:

1. Konversi data menjadi Dataframe.
2. Lakukan preprocessing data.
3. Hilangkan kolom 'CustomerID' dan 'gender'.
4. Latih model K-Means.
5. Buat plot untuk Elbow dan Cluster.

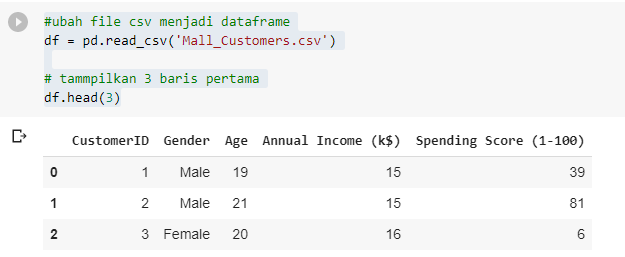
**Codelab**

Dataset yang akan Anda gunakan adalah data pengunjung sebuah mall fiktif. Dataset bisa Anda dapatkan pada [tautan](https://www.kaggle.com/vjchoudhary7/customer-segmentation-tutorial-in-python) berikut.

Setelah Anda berhasil mengunduh datanya, buka Notebook pada Google Colab dan unggah dataset Mall Customer Segmentation Data pada*session storage* Google Colab.

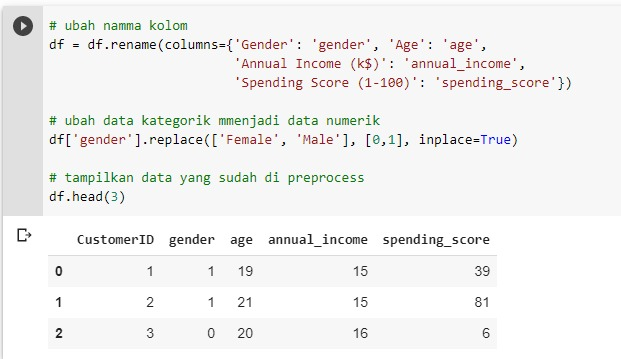
Pada cell pertama, kita ubah file csv kita ke dalam dataframe pandas dan menampilkan 3 baris pertama dari dataframe.

1. import pandas as pd
3. #ubah file csv menjadi dataframe
4. df = pd.read\_csv('Mall\_Customers.csv')
6. # tampilkan 3 baris pertama
7. df.head(3)

Tampilan dari 3 baris pertama dataframe di atas seperti berikut.  


Kemudian kita akan melakukan sedikit preprocessing yaitu mengubah nama kolom agar lebih seragam. Lalu kolom gender adalah kolom kategorik, maka kita akan mengubah data tersebut menjadi data numerik.

1. # ubah nama kolom
2. df = df.rename(columns={'Gender': 'gender', 'Age': 'age',
3. 'Annual Income (k$)': 'annual\_income',
4. 'Spending Score (1-100)': 'spending\_score'})
6. # ubah data kategorik menjadi data numerik
7. df['gender'].replace(['Female', 'Male'], [0,1], inplace=True)
9. # tampilkan data yang sudah di preprocess
10. df.head(3)

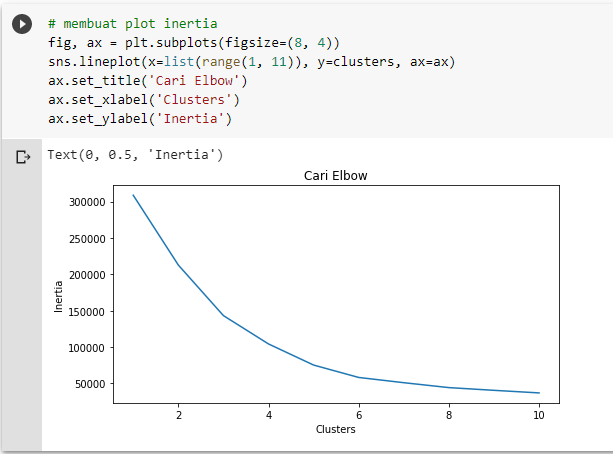
Setelah dilakukan preprocessing dengan mengubah nama kolom supaya lebih seragam, maka hasilnya seperti di bawah ini.  


Di tahap selanjutnya kita akan mengimpor K-Means. Di tahap ini juga kita akan menghilangkan kolom Customer ID dan gender karena kurang relevan untuk proses clustering. Selanjutnya kita akan menentukan nilai K yang optimal dengan metode Elbow. Library K-means dari SKLearn menyediakan fungsi untuk menghitung inersia dari K-Means dengan jumlah K tertentu. Di sini kita akan membuat list yang berisi inersia dari nilai K antara 1 sampai 11.

1. from sklearn.cluster import KMeans
3. # menghilangkan kolom customer id dan gender
4. X = df.drop(['CustomerID', 'gender'], axis=1)
6. #membuat list yang berisi inertia
7. clusters = []
8. for i in range(1,11):
9. km = KMeans(n\_clusters=i).fit(X)
10. clusters.append(km.inertia\_)

Jalankan kode di bawah untuk membuat plot inersia dari setiap nilai K. Sesuai plot di bawah, kita bisa melihat bahwa elbow berada di nilai K sama dengan 5, di mana penurunan inersia tidak lagi signifikan setelah nilai K sama dengan 5. Jangan lupa mengimpor library yang dibutuhkan untuk membuat plot ya.

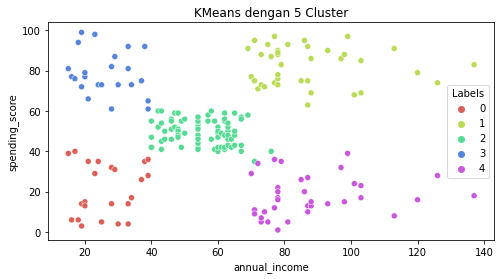
1. import matplotlib.pyplot as plt
2. %matplotlib inline
3. import seaborn as sns
5. # membuat plot inertia
6. fig, ax = plt.subplots(figsize=(8, 4))
7. sns.lineplot(x=list(range(1, 11)), y=clusters, ax=ax)
8. ax.set\_title('Cari Elbow')
9. ax.set\_xlabel('Clusters')
10. ax.set\_ylabel('Inertia')

Hasil dari kode di atas menampilkan plot inersia sebagai berikut.  


Terakhir kita bisa melatih kembali K-Means dengan jumlah K yang didapat dari metode Elbow. Lalu kita bisa membuat plot hasil pengklasteran K-Means dengan menjalankan kode di bawah.

1. # membuat objek KMeans
2. km5 = KMeans(n\_clusters=5).fit(X)
4. # menambahkan kolom label pada dataset
5. X['Labels'] = km5.labels\_
7. # membuat plot KMeans dengan 5 klaster
8. plt.figure(figsize=(8,4))
9. sns.scatterplot(x=X['annual\_income'], y=X['spending\_score'], hue=X['Labels'],
10. palette=sns.color\_palette('hls', 5))
11. plt.title('KMeans dengan 5 Cluster')
12. plt.show()

Sehingga jika kode di atas dijalankan, maka tampilan KMeans dengan 5 klaster seperti di bawah ini.

[](https://www.dicoding.com/academies/184/tutorials/8422?from=8417)