ibimbing

BOOT CAMP

DATA SCIENCE

Kaylila Larasati

ibimbing

Menggunakan 3 library utama:

- pandas: untuk manipulasi data
- seaborn: untuk visualisasi statistik
- matplotlib: untuk menampilkan grafik

- Dataset diambil dari file Mall_Customer.csv.
- Data dimuat menggunakan pandas, dan diperiksa ukuran serta beberapa data awal (head).
- Ukuran data: misalnya (200, 5) berarti ada 200 baris dan 5 kolom.

```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv('/content/Mall_Customers.csv')
df
```

	CustomerID	Genre	Age	Annual_Income_(k$)	Spending_Score
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40
...	...	...	...	...	...
195	196	Female	35	120	79
196	197	Female	45	126	28
197	198	Male	32	126	74
198	199	Male	32	137	18
199	200	Male	30	137	83

200 rows × 5 columns

Cek Info & Missing Value

- df.info() digunakan untuk melihat tipe data dan jumlah non-null.
- df.isnull().sum() untuk mengecek apakah ada nilai kosong (missing value) di kolom manapun.

```
print("\nInfo Dataset:")
print(df.info())
```

```
Info Dataset:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 200 entries, 0 to 199
Data columns (total 5 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   CustomerID              200 non-null    int64
 1   Genre                   200 non-null    object
 2   Age                     200 non-null    int64
 3   Annual_Income_(k$)      200 non-null    int64
 4   Spending_Score          200 non-null    int64
dtypes: int64(4), object(1)
memory usage: 7.9+ KB
None
```

```
print("\nCek Missing Value:")
print(df.isnull().sum())
```

HANDLING MISSING VALUE

- Jika ada nilai kosong di kolom numerik seperti Age, bisa diisi dengan median menggunakan fillna().
- Tidak semua kolom punya missing value, jadi hanya dilakukan jika diperlukan.

```
print("\nJumlah Data Duplikat:", df.duplicated().sum())
```

```
Jumlah Data Duplikat: 0
```

CEK & HAPUS DUPLIKAT

- df.duplicated().sum() mengecek apakah ada baris data yang sama persis (duplikat).
- Jika ada, baris duplikat dihapus agar tidak mengganggu analisis.

STATISTIK DESKRIPTIF

- df.describe() memberikan statistik dasar seperti mean, min, max, dan quartile untuk kolom numerik.
- Berguna untuk memahami rentang dan sebaran data.

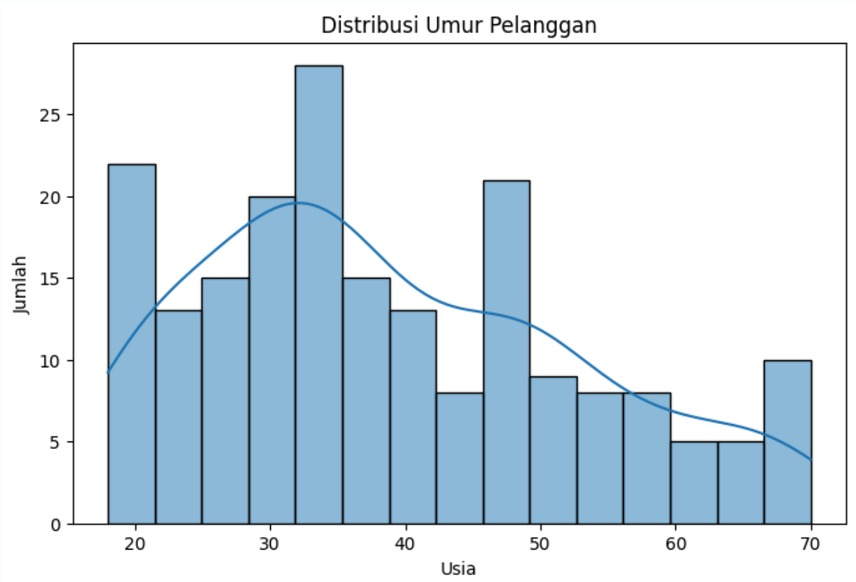
```
print("\nStatistik Deskriptif:")
print(df.describe())
```

```
Statistik Deskriptif:
       CustomerID         Age  Annual_Income_(k$)  Spending_Score
count  200.000000  200.000000          200.000000      200.000000
mean   100.500000   38.850000           60.560000       50.200000
std     57.879185   13.969007           26.264721       25.823522
min      1.000000   18.000000           15.000000        1.000000
25%     50.750000   28.750000           41.500000       34.750000
50%    100.500000   36.000000           61.500000       50.000000
75%    150.250000   49.000000           78.000000       73.000000
max    200.000000   70.000000          137.000000       99.000000
```

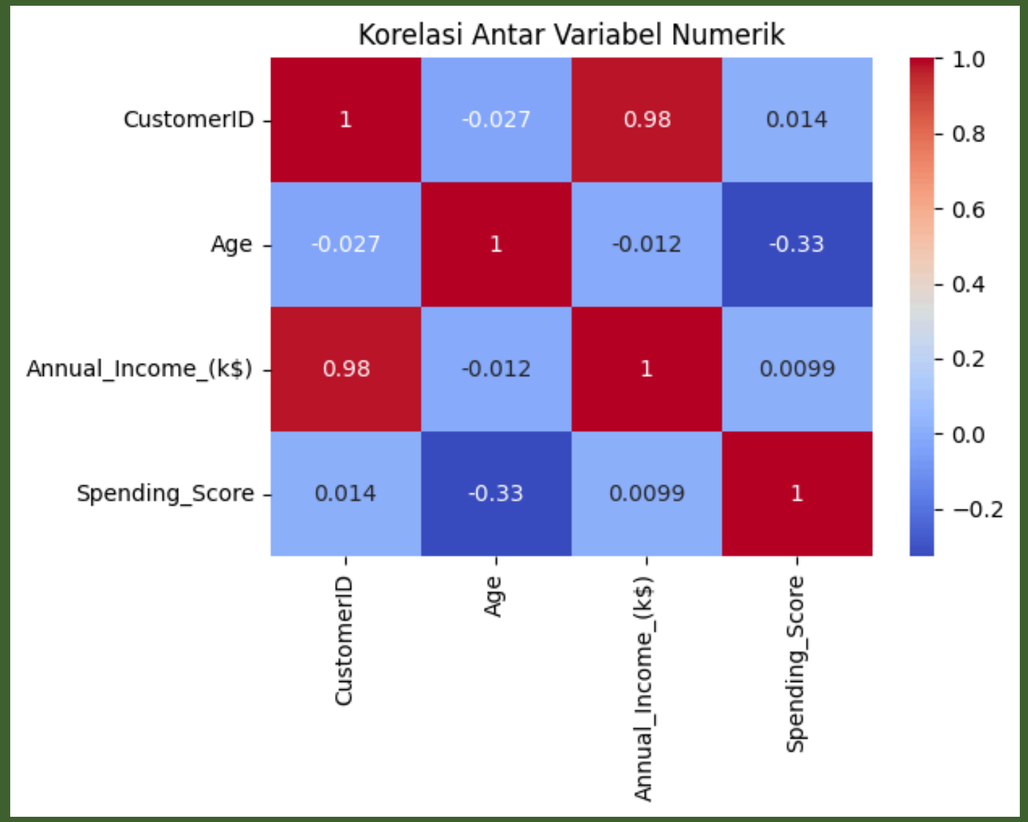
Visualisasi Distribusi Umur Pelanggan

Kode ini digunakan untuk menampilkan distribusi usia pelanggan dalam bentuk histogram. Dengan bantuan sns.histplot(), kita dapat melihat rentang usia yang paling dominan dalam dataset, serta mengetahui apakah sebaran usia pelanggan merata atau terpusat pada kelompok tertentu.

```
plt.figure(figsize=(8,5))
sns.histplot(df['Age'], kde=True, bins=15)
plt.title('Distribusi Umur Pelanggan')
plt.xlabel('Usia')
plt.ylabel('Jumlah')
plt.show()
```



```
plt.figure(figsize=(6,4))
sns.heatmap(df.corr(numeric_only=True), annot=True, cmap='coolwarm')
plt.title('Korelasi Antar Variabel Numerik')
plt.show()
```



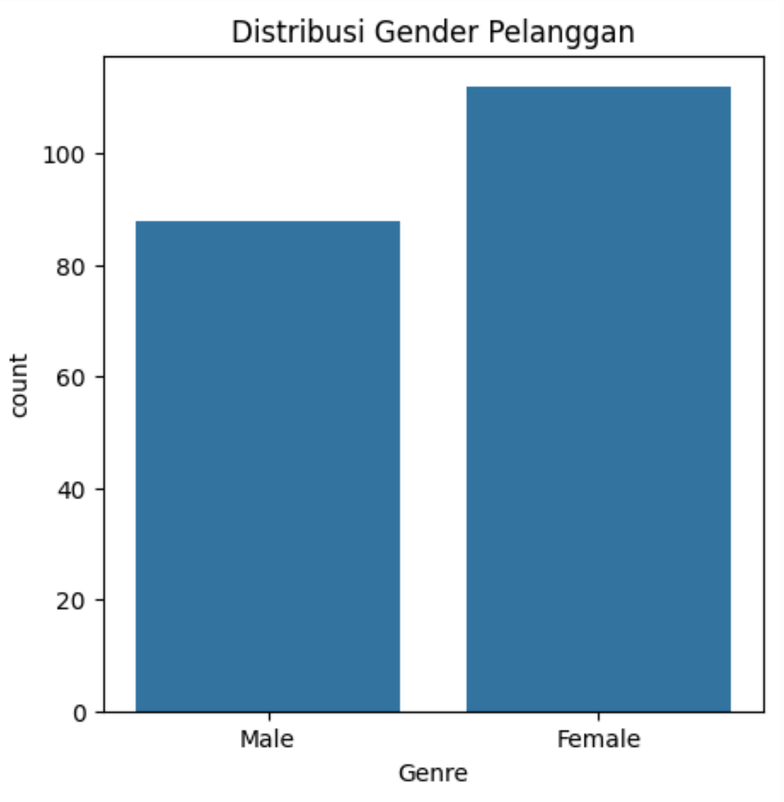
HEATMAP KORELASI

- Menggunakan fungsi sns.heatmap() dari Seaborn.
- Menampilkan hubungan antar variabel numerik: Age, Annual Income, dan Spending Score.
- Nilai korelasi ditampilkan dalam bentuk angka dan warna (biru ke merah).
- Membantu melihat apakah antar variabel memiliki hubungan positif, negatif, atau tidak berkorelasi.

Distribusi Gender Pelanggan

Kode ini memanfaatkan sns.countplot() untuk menunjukkan jumlah pelanggan berdasarkan gender. Grafik ini memberikan gambaran proporsi antara pelanggan pria dan wanita dalam dataset, sehingga dapat dianalisis apakah ada dominasi gender tertentu dalam data pelanggan tersebut.

```
plt.figure(figsize=(5,5))
sns.countplot(data=df, x='Genre')
plt.title('Distribusi Gender Pelanggan')
plt.show()
```



TERIMA KASIH

ibimbing