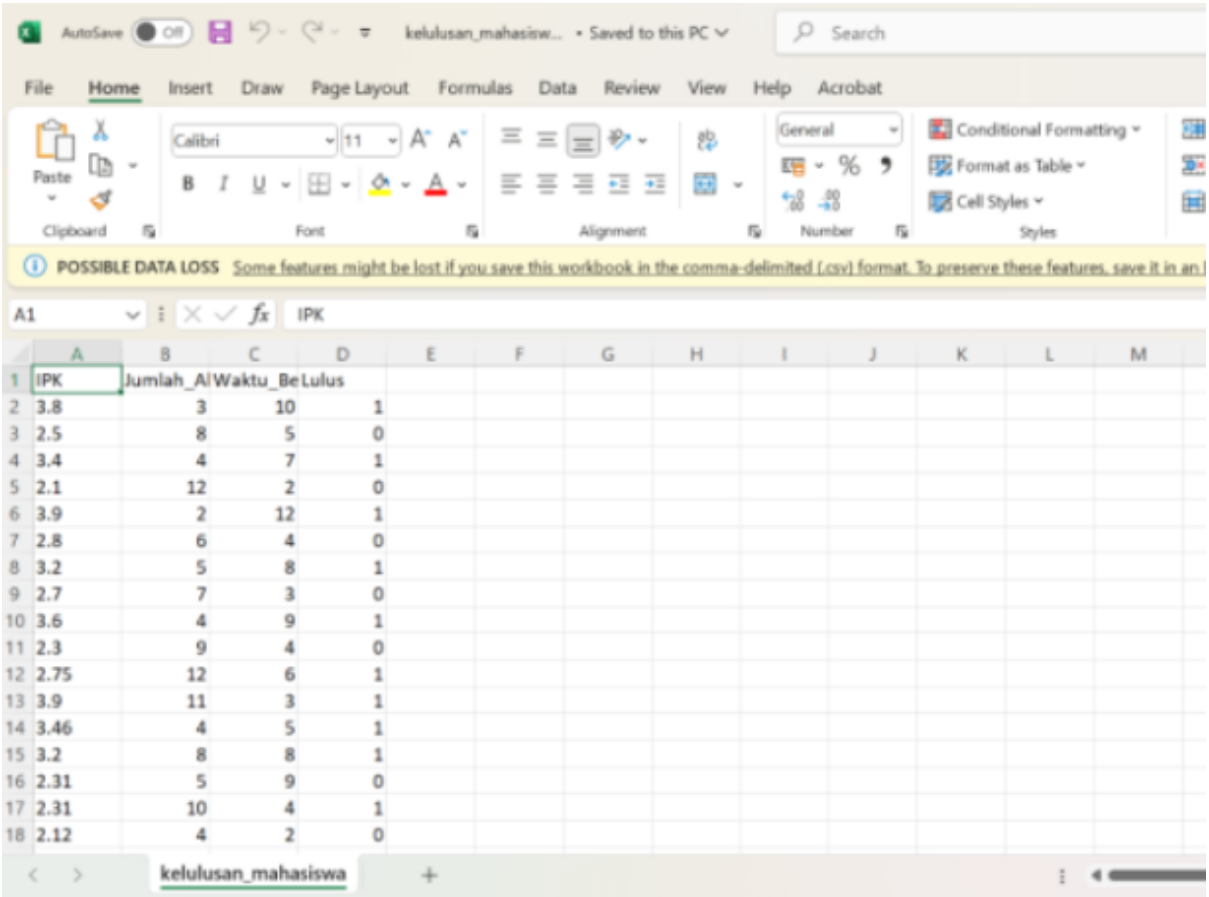
Nama : Zulyan Widyaka Krisna
NIM : 231011403446
Kelas : 05TPLE016

PERTEMUAN 4

1. Persiapan Data

Persiapkan data dalam bentuk tabel menggunakan Microsoft Excel. Pastikan semua data sudah tersusun dengan rapi dan lengkap agar mudah diolah. Setelah itu, simpan file tersebut dalam format **Comma Separated Values (.csv)**. Format ini sangat disarankan karena lebih kompatibel dengan bahasa pemrograman Python dan memudahkan proses analisis data menggunakan library seperti *pandas*.

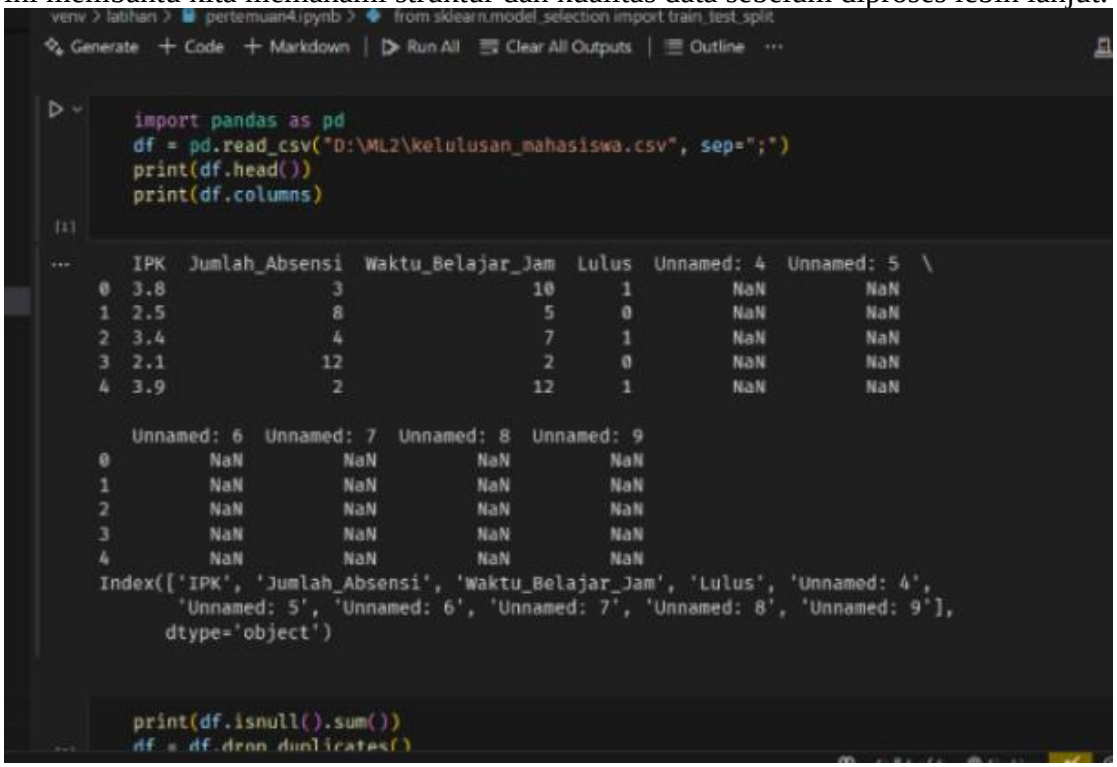


	A	B	C	D
1	IPK	Jumlah_Al	Waktu_Be	Lulus
2	3.8	3	10	1
3	2.5	8	5	0
4	3.4	4	7	1
5	2.1	12	2	0
6	3.9	2	12	1
7	2.8	6	4	0
8	3.2	5	8	1
9	2.7	7	3	0
10	3.6	4	9	1
11	2.3	9	4	0
12	2.75	12	6	1
13	3.9	11	3	1
14	3.46	4	5	1
15	3.2	8	8	1
16	2.31	5	9	0
17	2.31	10	4	1
18	2.12	4	2	0

2. Membaca Data dengan Pandas

Setelah file CSV tersedia, buka file tersebut di Python menggunakan library **pandas**. Tahapan ini penting untuk memastikan data berhasil dimuat dengan benar. Selanjutnya, tampilkan informasi dasar mengenai dataset, seperti jumlah baris dan kolom, tipe data pada tiap atribut, serta periksa apakah terdapat data kosong (missing values) atau tidak. Langkah

ini membantu kita memahami struktur dan kualitas data sebelum diproses lebih lanjut.



```
import pandas as pd
df = pd.read_csv("D:\ML2\kelulusan_mahasiswa.csv", sep=";")
print(df.head())
print(df.columns)
```

```
   IPK  Jumlah_Absensi  Waktu_Belajar_Jam  Lulus  Unnamed: 4  Unnamed: 5  \
0  3.8               3                 10      1         NaN         NaN
1  2.5               8                  5      0         NaN         NaN
2  3.4               4                  7      1         NaN         NaN
3  2.1              12                  2      0         NaN         NaN
4  3.9               2                 12      1         NaN         NaN

   Unnamed: 6  Unnamed: 7  Unnamed: 8  Unnamed: 9
0         NaN         NaN         NaN         NaN
1         NaN         NaN         NaN         NaN
2         NaN         NaN         NaN         NaN
3         NaN         NaN         NaN         NaN
4         NaN         NaN         NaN         NaN
Index(['IPK', 'Jumlah_Absensi', 'Waktu_Belajar_Jam', 'Lulus', 'Unnamed: 4',
       'Unnamed: 5', 'Unnamed: 6', 'Unnamed: 7', 'Unnamed: 8', 'Unnamed: 9'],
      dtype='object')
```

```
print(df.isnull().sum())
df = df.drop_duplicates()
```

3. Pengecekan dan Penanganan Data Hilang

Lakukan pemeriksaan terhadap data yang hilang pada setiap kolom. Jika ditemukan nilai kosong, lakukan penanganan dengan cara:

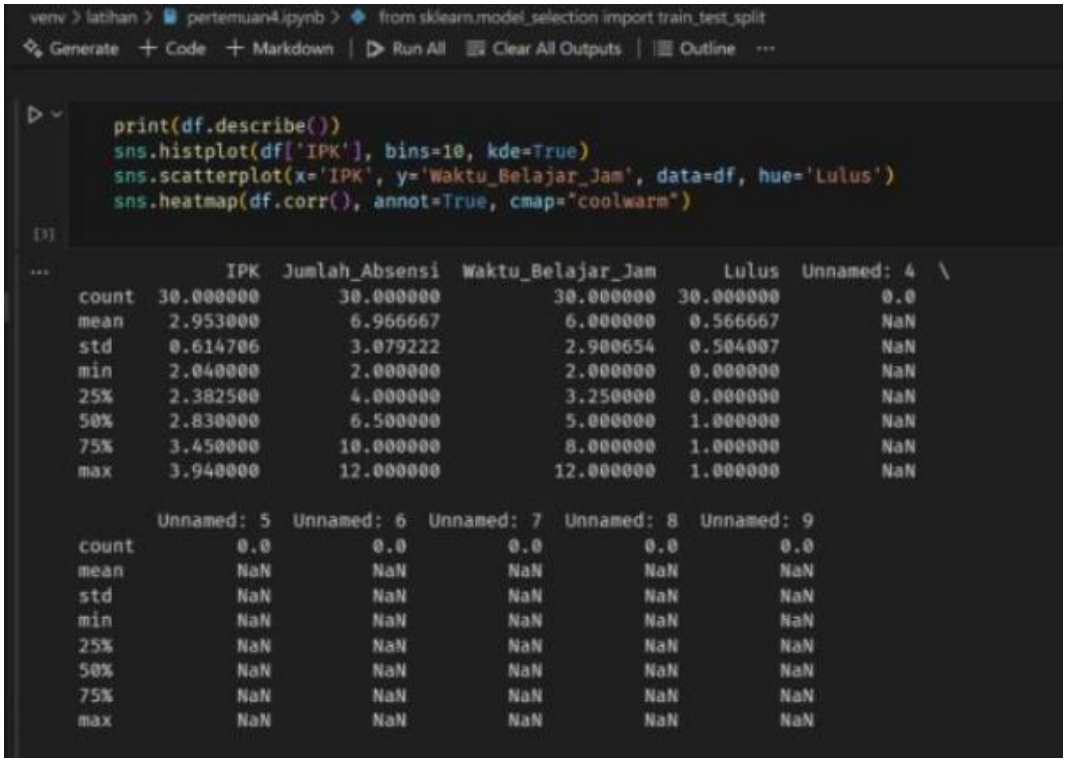
- Mengisi nilai kosong menggunakan **median** untuk data numerik
- Menggunakan **modus** untuk data kategorikal

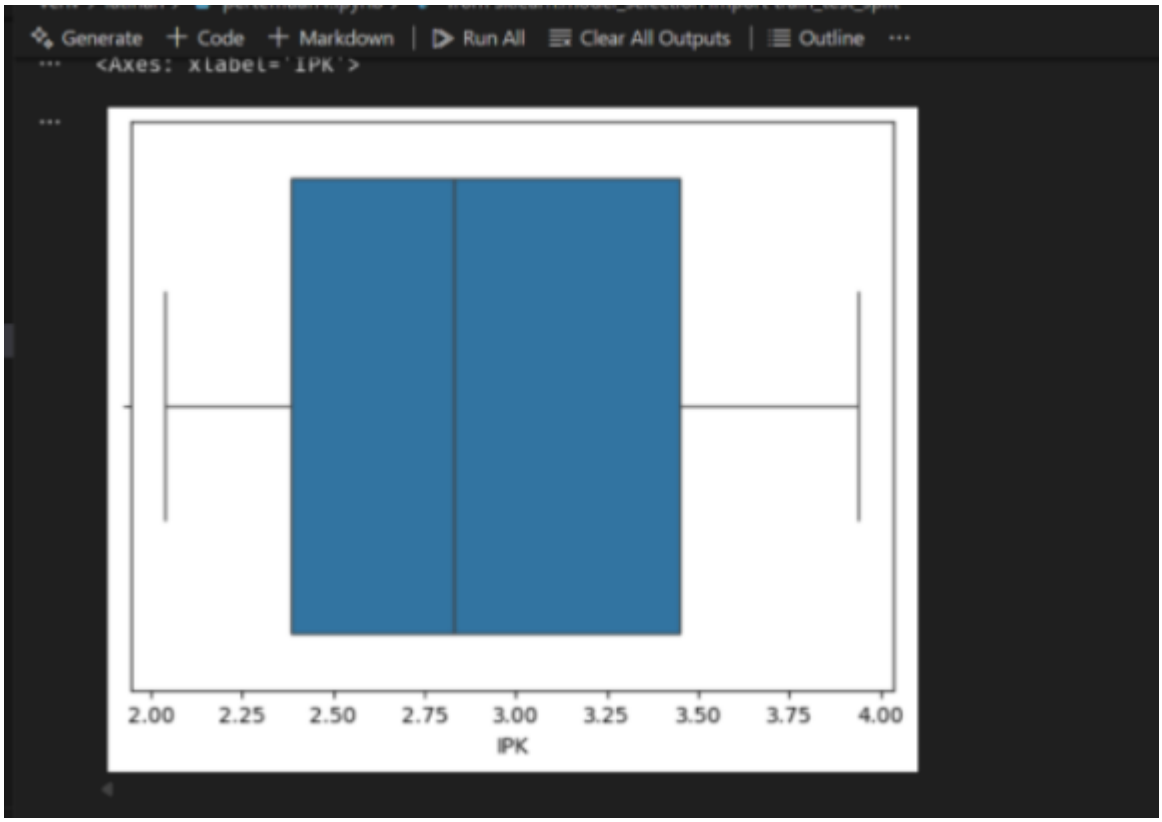
Selain itu, hapus data yang duplikat agar hasil analisis tidak bias. Untuk mendeteksi adanya **outlier**, gunakan visualisasi **boxplot**. Grafik ini membantu kita mengidentifikasi data yang berada jauh dari rentang sebaran normalnya.

4. Analisis Statistik dan Visualisasi Data

Lakukan analisis statistik deskriptif untuk memahami karakteristik data, seperti:

- Nilai rata-rata (mean)
- Median
- Standar deviasi
- Nilai minimum dan maksimum
- Kuartil

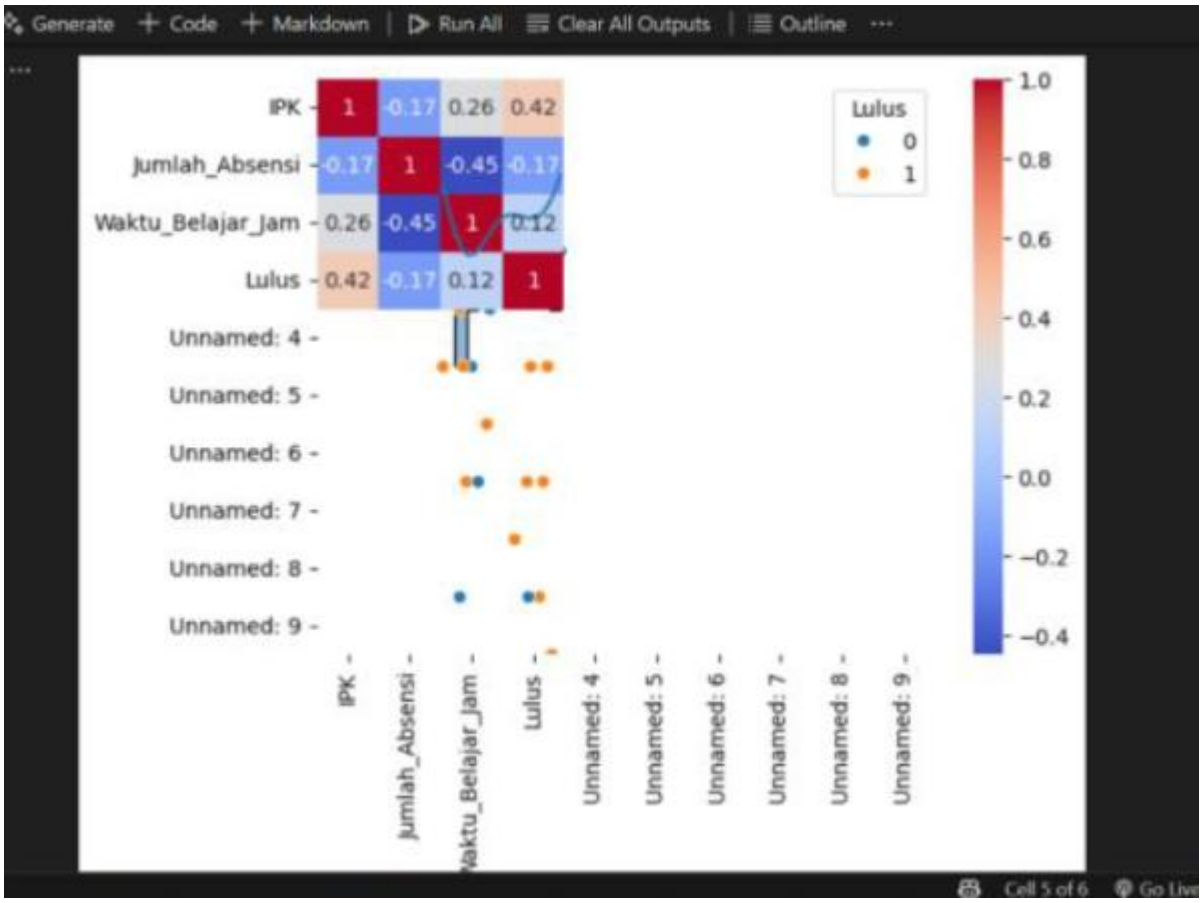




Gunakan **histogram** untuk melihat distribusi nilai IPK, serta **scatterplot** untuk memvisualisasikan hubungan antara IPK dan waktu belajar. Untuk melihat korelasi antarvariabel, tampilkan **heatmap** agar pola hubungan antar fitur terlihat lebih jelas dan informatif.

5. Pembuatan Fitur Turunan (*Feature Engineering*)

Tambahkan variabel baru yang relevan guna meningkatkan kualitas analisis maupun performa model prediksi di tahap selanjutnya. Proses ini dikenal dengan istilah *feature engineering* dan sangat penting untuk membuat model menjadi lebih akurat dan kontekstual.



6. Pembagian Dataset

Tahap terakhir adalah membagi dataset menjadi tiga bagian:

- **Data pelatihan (train)** sebesar 70%
- **Data validasi (validation)** sebesar 15%
- **Data pengujian (test)** sebesar 15%

Gunakan metode **stratified split** agar proporsi kelas di setiap subset tetap seimbang dan representatif terhadap keseluruhan dataset. Langkah ini memastikan model yang dibangun nantinya tidak bias terhadap kelas tertentu dan memiliki kemampuan generalisasi yang baik.