JOBSHEET 3

**PENGENALAN PYTHON DALAM STATISTIKA, DISTRIBUSI NORMAL DAN IMPUTASI DATA**

**Nama : Lovie Jechonia Tonimba**

**NIM : 244107060101**

**Class : 2G**

**Major : Business Information System**

TUJUAN

1. Mahasiswa mampu melakukan operasi statistika dasar menggunakan python
2. Mahasiswa dapat melakukan imputasi data jika ada data yang kosong

# PENJELASAN UMUM

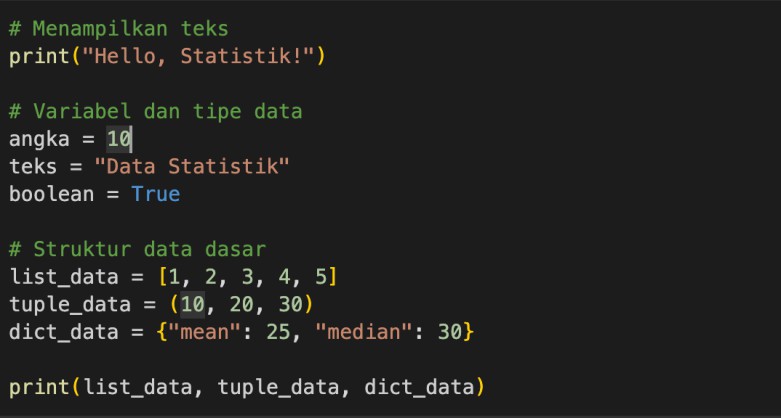
* + Pada ujicoba ini, menggunakan data ecommerce\_consumer\_behaviour yang dapat diunduh di sini:

[https://drive.google.com/file/d/1gNBfaLOm-u17kSQew-](https://drive.google.com/file/d/1gNBfaLOm-u17kSQew-SLBmBiQ1EgBijd/view?usp=drive_link) [SLBmBiQ1EgBijd/view?usp=drive\_link](https://drive.google.com/file/d/1gNBfaLOm-u17kSQew-SLBmBiQ1EgBijd/view?usp=drive_link)

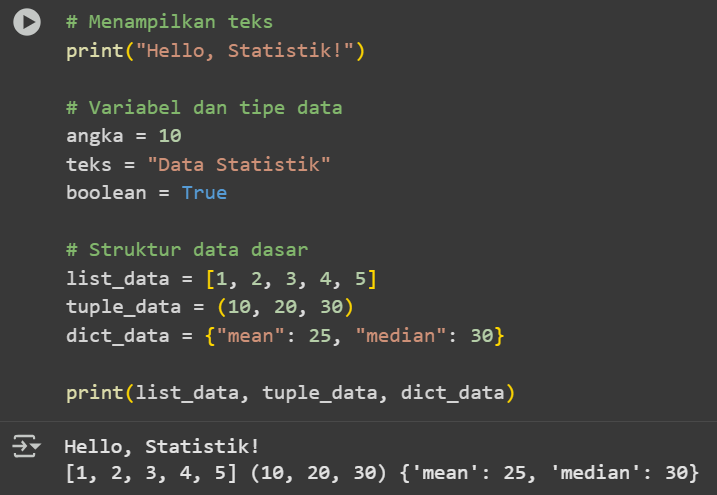
* + Untuk menjalankan kode python, dapat menggunakan Google Collaboratory atau Visual Studio Code, dengan menginstal python sebelumnya

# Bagian 1: Perintah Dasar

Sebelum melakukan analisis statistik, kita perlu memahami beberapa perintah dasar dalam Python.



Silakan dicoba, bagaimana hasilnya?

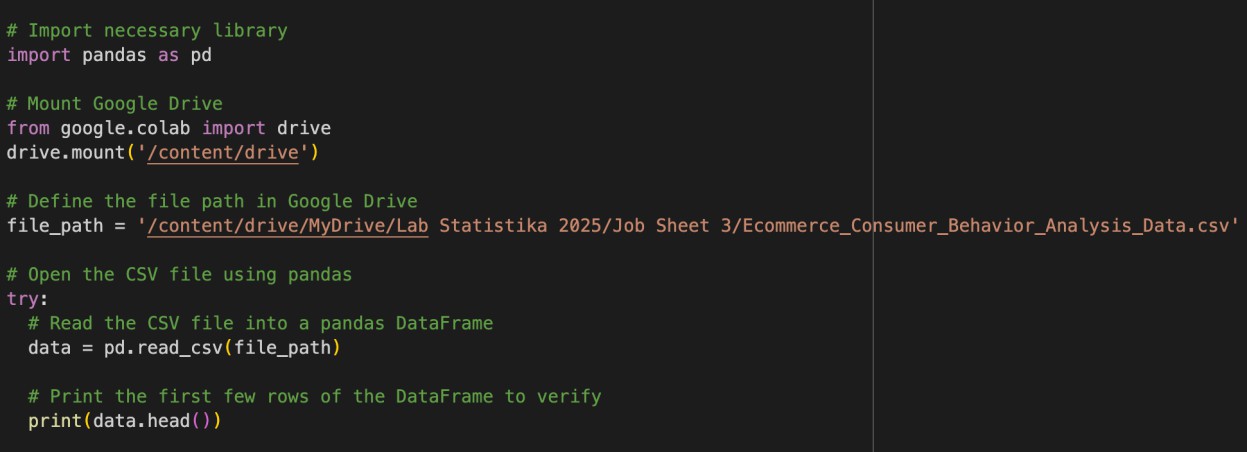


Perintah print("Hello, Statistik!") digunakan untuk menampilkan teks ke layar. Setelah itu, dibuat beberapa variabel dengan tipe data berbeda, yaitu angka (integer), teks (string), dan boolean (nilai True/False).

* List (list\_data = [1, 2, 3, 4, 5]) → kumpulan data yang bisa diubah (mutable).
* Tuple (tuple\_data = (10, 20, 30)) → kumpulan data tetap (immutable).
* Dictionary (dict\_data = {"mean": 25, "median": 30}) → pasangan *key-value* untuk menyimpan data dengan label.

Kemudian perintah print(list\_data, tuple\_data, dict\_data) dijalankan, Python menampilkan semua nilai tersebut di bawahnya

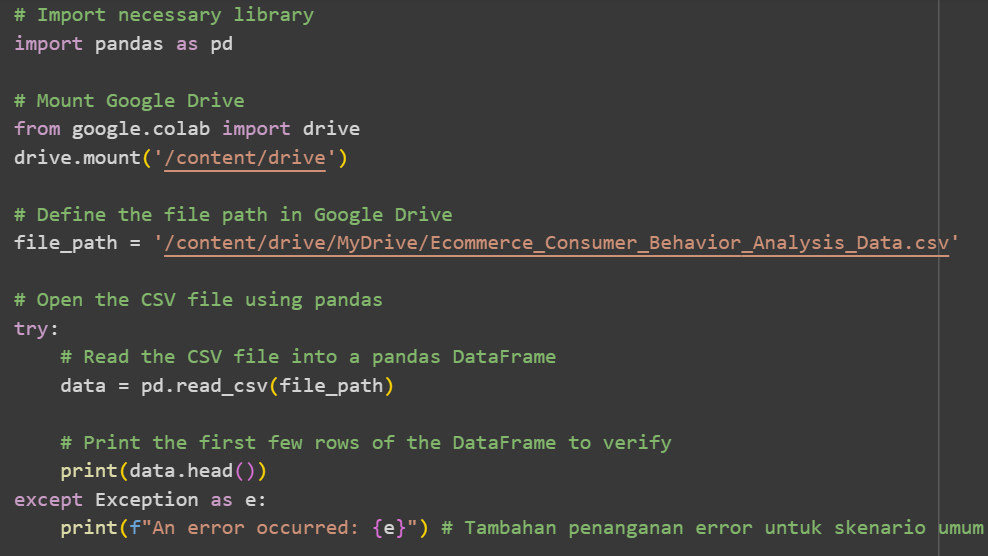
# Bagian 2: Membuka Dataset

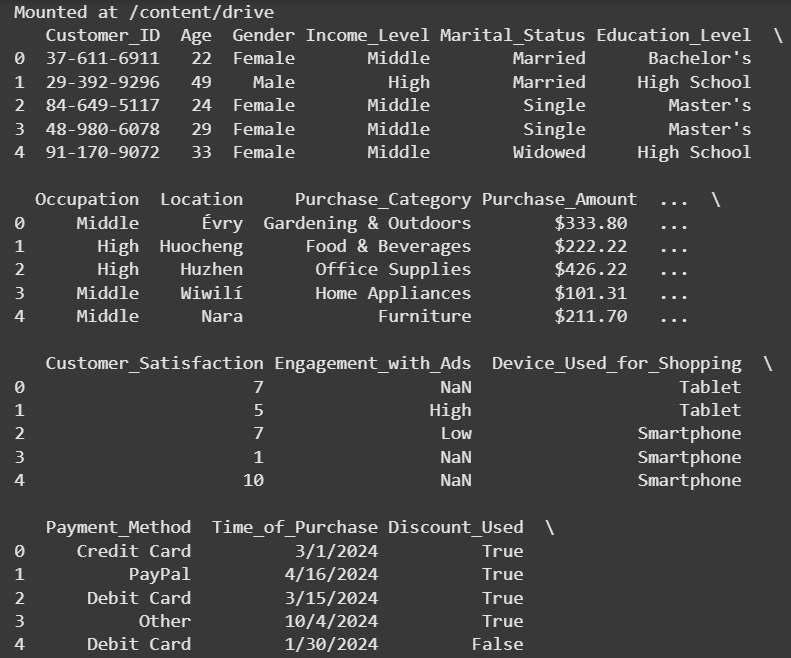
Setelah mengetahui perintah-perintah dasar dalam python, selanjutnya yang perlu dipelajari adalah cara membuka dataset.

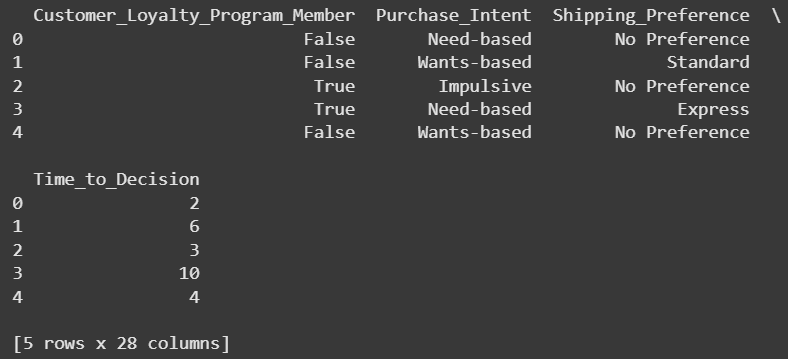
Catatan:

Pada contoh di atas, data diambil dari Google Drive dengan lokasi file yang sama dengan file python. Untuk cara membuka data yang lain, silakan dicari.

Silakan dicoba dan bagaimana hasilnya?



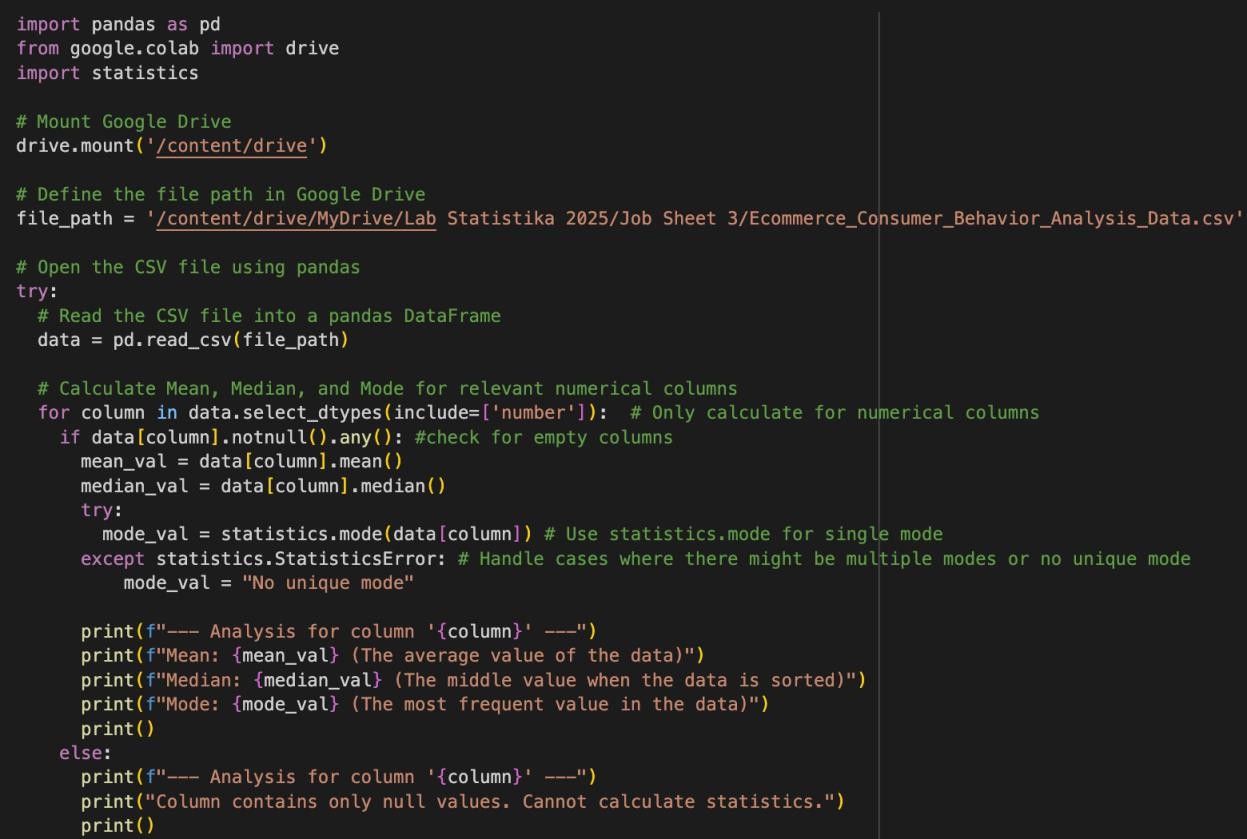




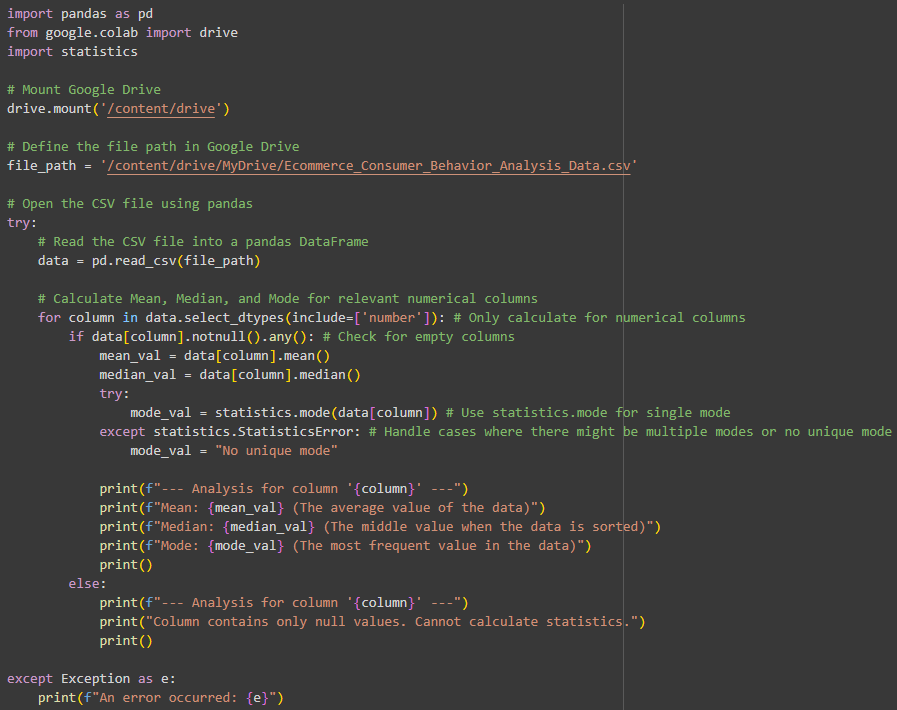
File dataset diambil dari Google Drive, kemudian dimuat ke Colab dengan perintah drive.mount('/content/drive'), sehingga Python dapat mengakses file yang tersimpan di Drive. Setelah itu, dataset dibuka dengan perintah seperti pd.read\_csv() dan ditampilkan menggunakan df.head(), yang menampilkan 5 baris pertama dari total 28 kolom.

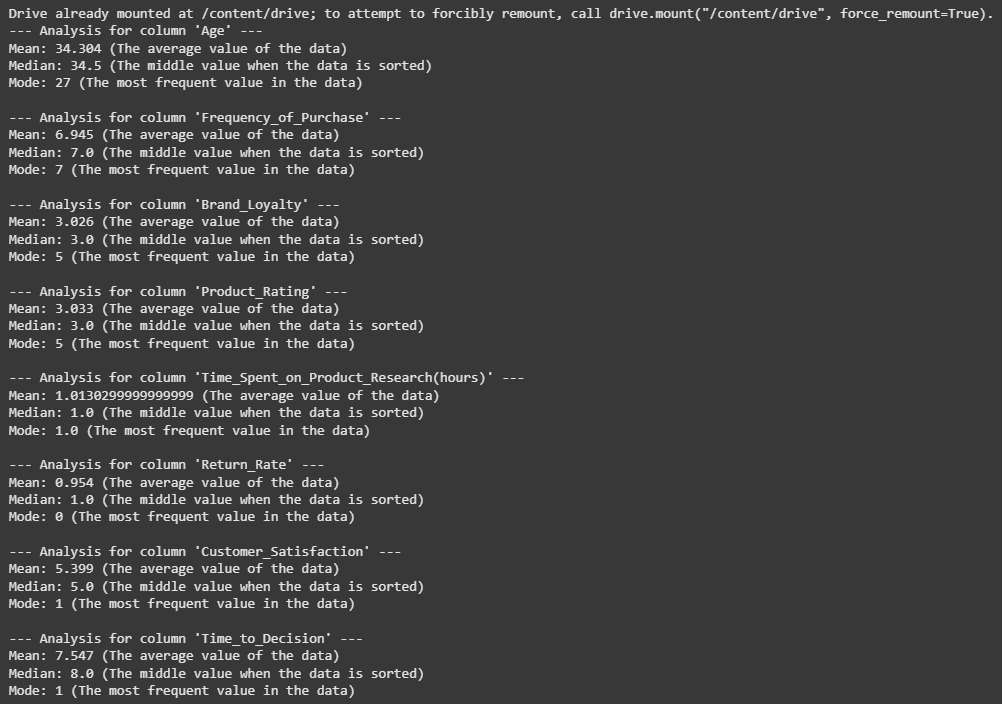
Dari hasil tampilan, dataset berisi berbagai informasi perilaku konsumen e-commerce, seperti Customer\_ID, Age, Gender, Income\_Level, Marital\_Status, Education\_Level, Purchase\_Amount, Payment\_Method, hingga Device\_Used\_for\_Shopping.  
Selain itu, juga terlihat adanya nilai NaN (Not a Number) pada beberapa kolom seperti *Device\_Used\_for\_Shopping* dan *Engagement\_with\_Ads*, yang menandakan adanya data kosong atau hilang (missing values) — hal ini nanti akan menjadi fokus pada bagian imputasi data.

# Bagian 3: Perhitungan Pemusatan Data

Pada bagian ini akan ditunjukkan bagaimana cara menghitung pemusatan data, yakni Mean, Median dan Modus, berdasarkan data yang ada.

Silakan dicoba, bagaimana hasilnya?



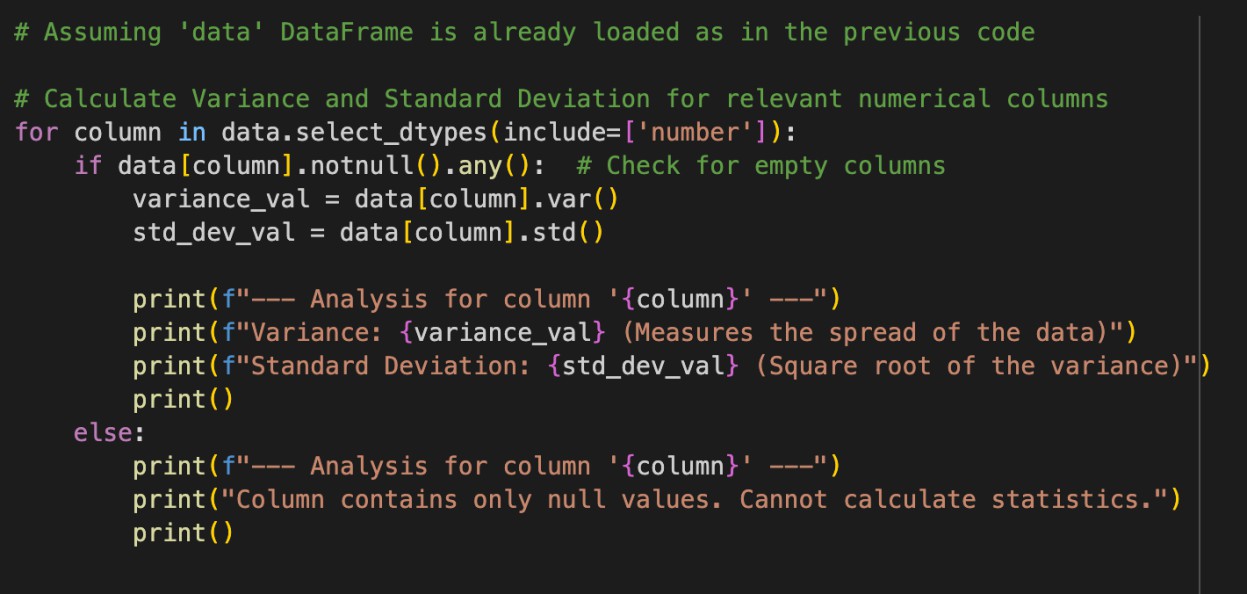


Program ini menghitung ukuran pemusatan data (central tendency) yaitu Mean (rata-rata), Median (nilai tengah), dan Mode (nilai yang paling sering muncul) untuk beberapa kolom numerik dalam dataset seperti Age, Frequency\_of\_Purchase, Brand\_Loyalty, Product\_Rating, Time\_Spent\_on\_Product\_Research(hours), Return\_Rate, Customer\_Satisfaction, dan Time\_to\_Decision.

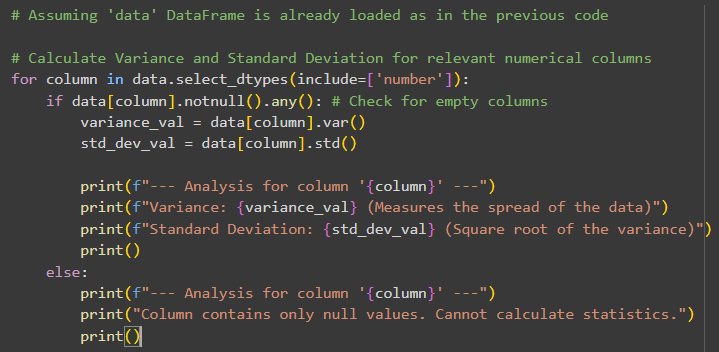
* Mean menunjukkan rata-rata nilai dari seluruh data. Misalnya pada kolom Age, nilai mean = 34.304 yang berarti rata-rata usia pelanggan sekitar 34 tahun.
* Median adalah nilai tengah dari data yang sudah diurutkan. Misalnya median pada Frequency\_of\_Purchase = 7 berarti separuh pelanggan melakukan pembelian lebih dari 7 kali, dan separuhnya kurang dari 7 kali.
* Mode adalah nilai yang paling sering muncul, misalnya Mode = 27 pada kolom Age berarti usia 27 tahun adalah yang paling sering muncul di data pelanggan.

Dari hasil keseluruhan, terlihat bahwa nilai mean, median, dan mode pada tiap kolom tidak berbeda jauh, yang mengindikasikan distribusi data relatif seimbang atau mendekati normal.

# Bagian 4: Perhitungan Variablitas Data

Pada bagian ini akan dilakukan penghitungan variabilitas data, seperti variance dan standart deviasi.

Silakan dicoba, bagaimana hasilnya?



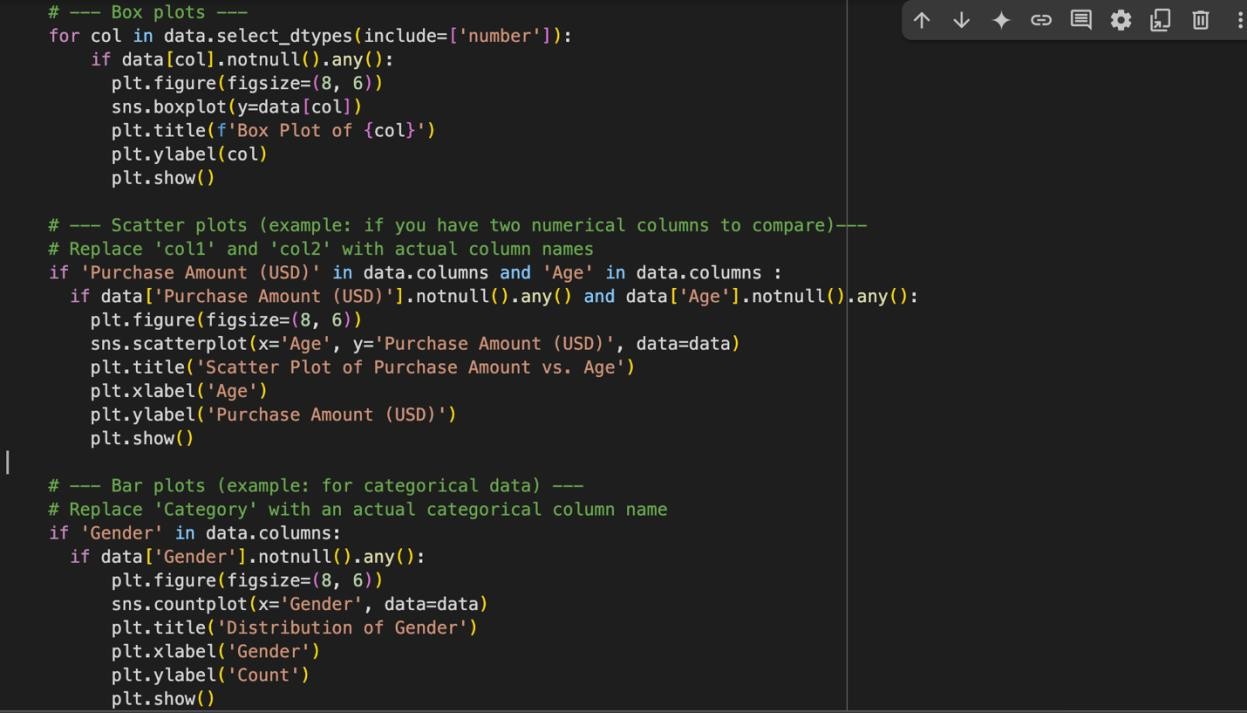
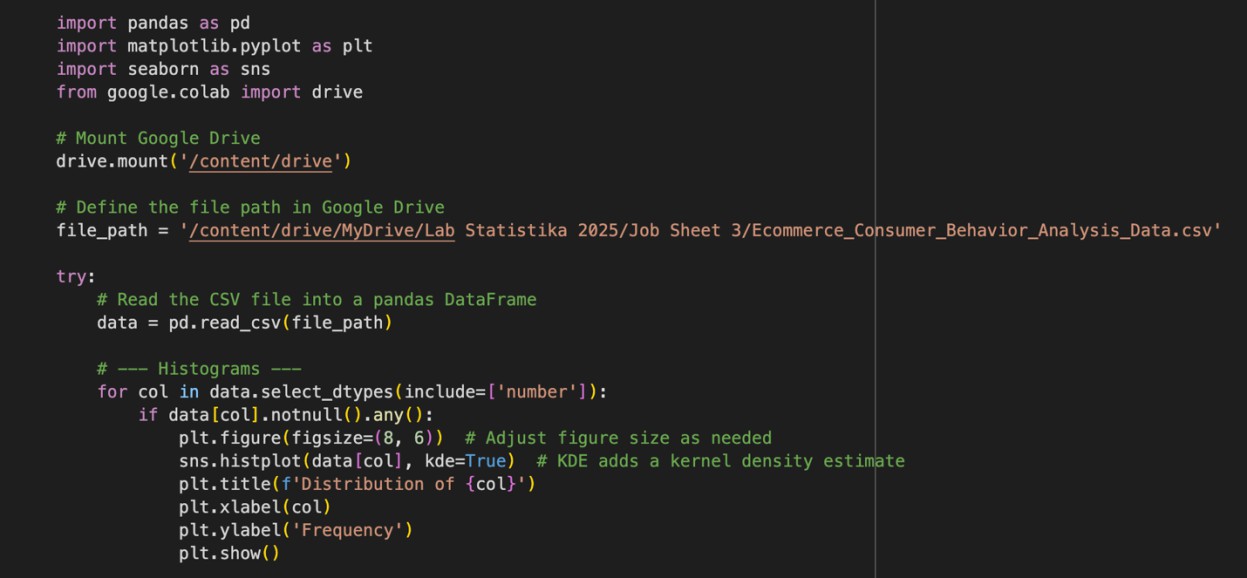


Program menghitung dua ukuran utama variabilitas data, yaitu variance (ragam) dan standard deviation (simpangan baku) untuk beberapa kolom numerik seperti Age, Frequency\_of\_Purchase, Brand\_Loyalty, Product\_Rating, Time\_Spent\_on\_Product\_Research(hours), Return\_Rate, Customer\_Satisfaction, dan Time\_to\_Decision.

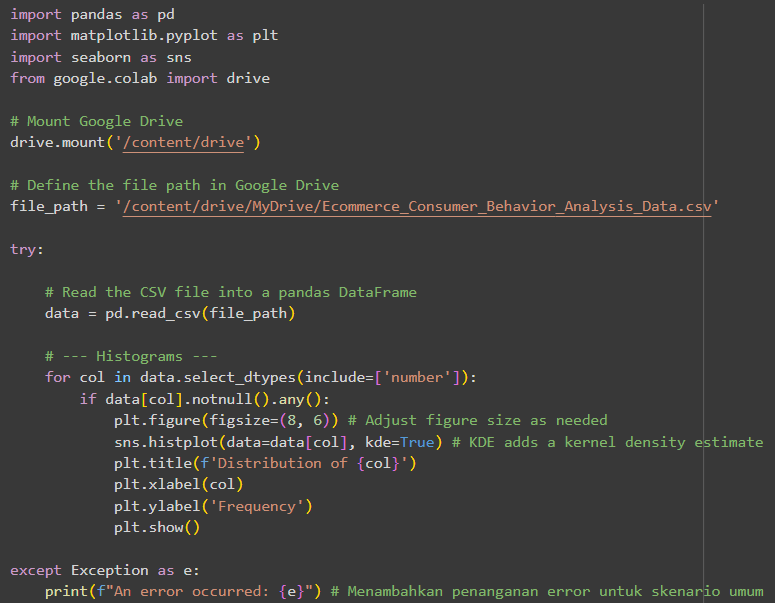
* Variance menggambarkan seberapa besar data menyebar dari rata-ratanya. Nilai variance yang besar berarti data lebih bervariasi atau tersebar luas.  
  Misalnya, pada kolom Age, variance = 87.48 yang menunjukkan penyebaran usia pelanggan cukup besar.
* Standard Deviation adalah akar kuadrat dari variance yang digunakan untuk menunjukkan seberapa jauh nilai-nilai data menyimpang dari rata-rata dalam satuan aslinya.  
  Contohnya, Standard Deviation Age = 9.35 berarti rata-rata perbedaan usia tiap pelanggan terhadap mean sekitar 9 tahun.

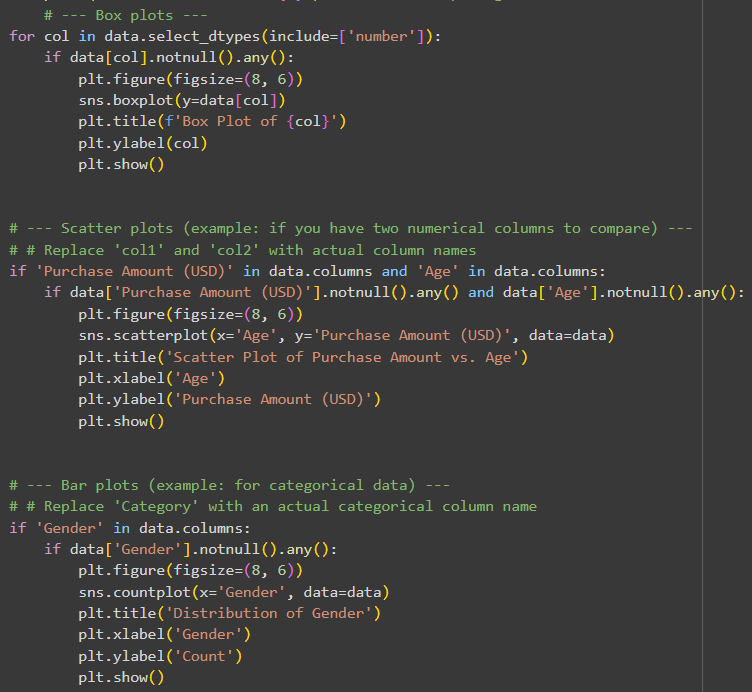
Dari hasil di atas, terlihat bahwa kolom seperti Age dan Time\_to\_Decision memiliki nilai variance dan standard deviation yang relatif tinggi, artinya data pada kolom tersebut lebih beragam atau tidak homogen. Sedangkan kolom seperti Return\_Rate dan Time\_Spent\_on\_Product\_Research(hours) memiliki nilai yang kecil, menunjukkan bahwa datanya lebih konsisten atau tidak banyak perbedaan antarobservasi.

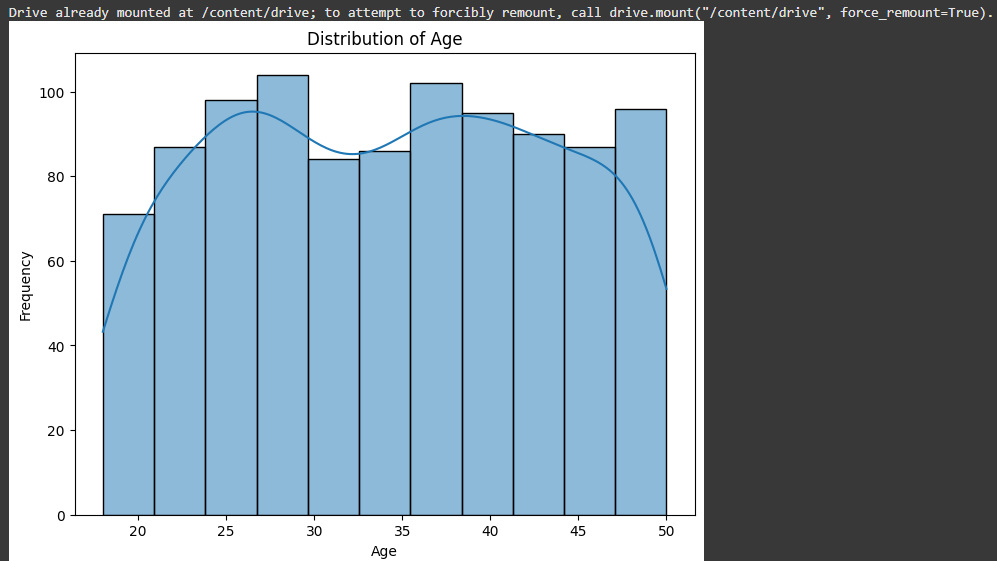
# Bagian 5: Penyajian Data

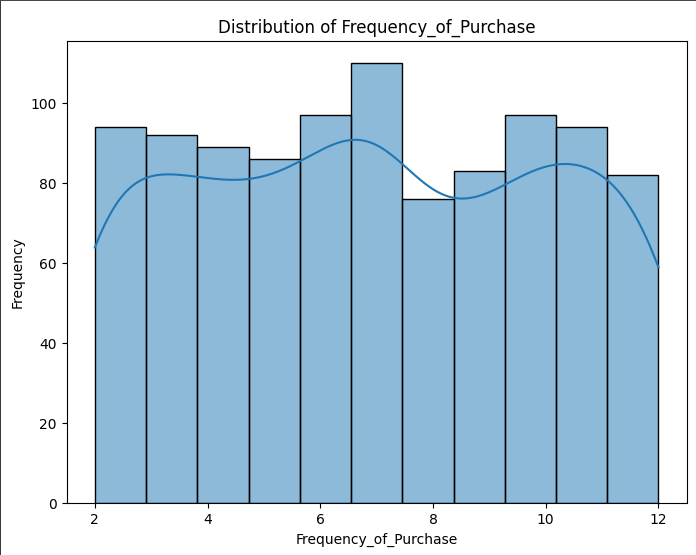
Pada bagian ini akan dicontohkan beberapa cara menyajikan data, menggunakan data ecommerce\_consumer\_behaviour

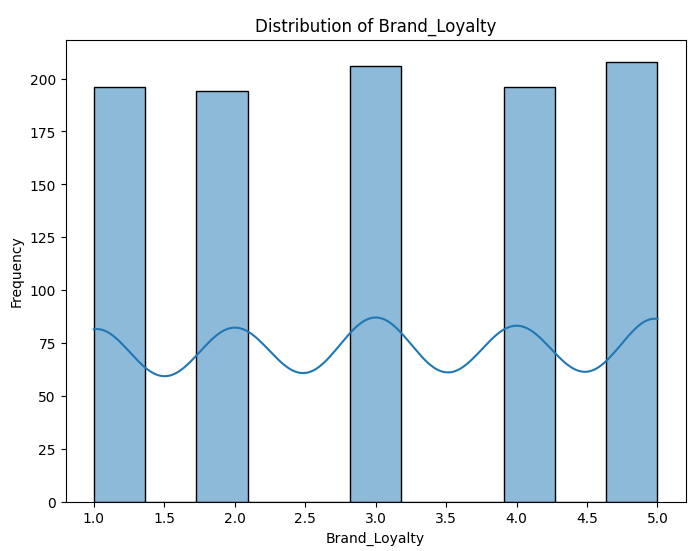
Silakan dicoba, bagaimana hasilnya?

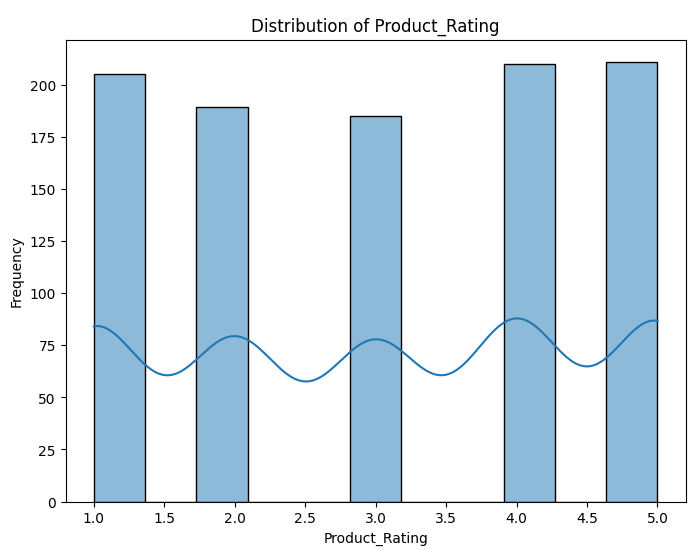


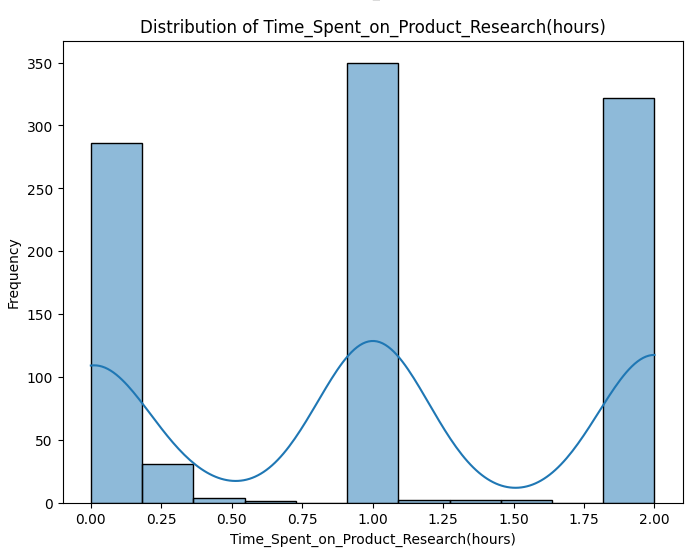


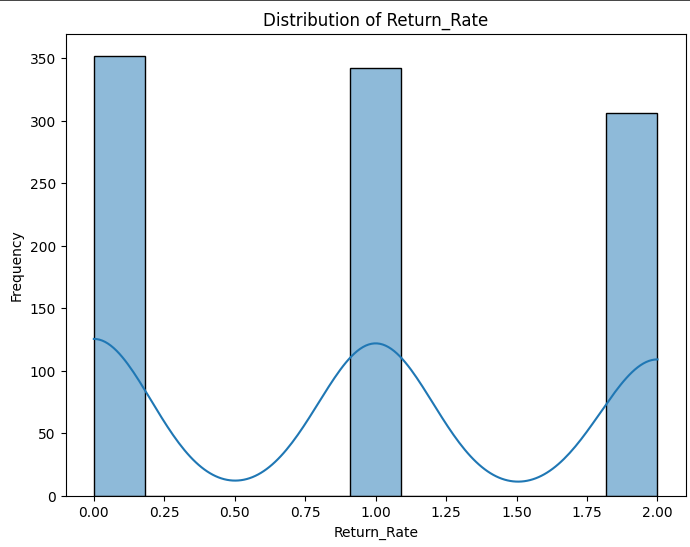


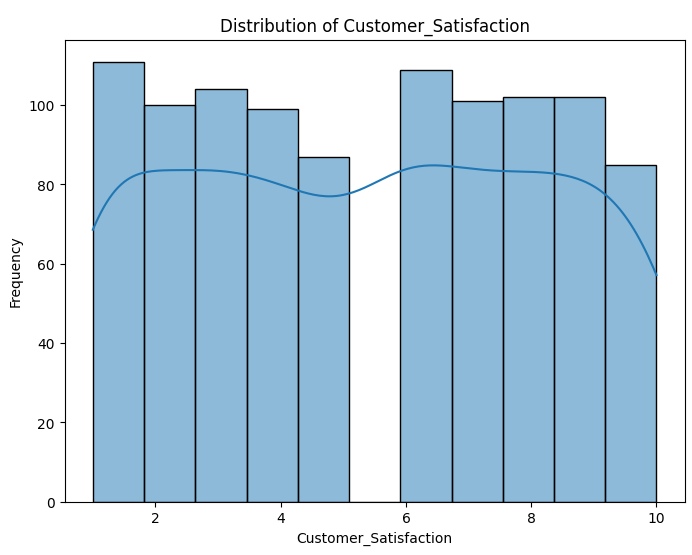


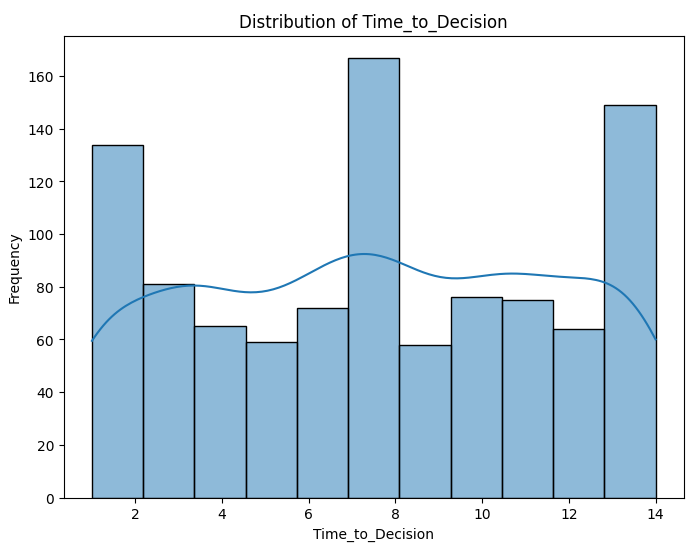


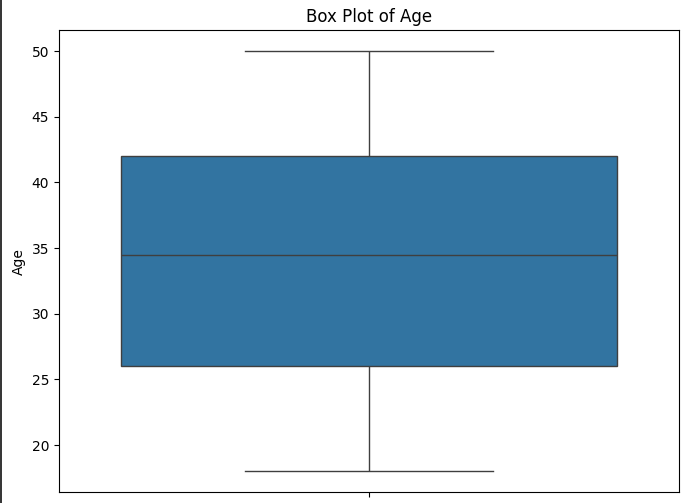


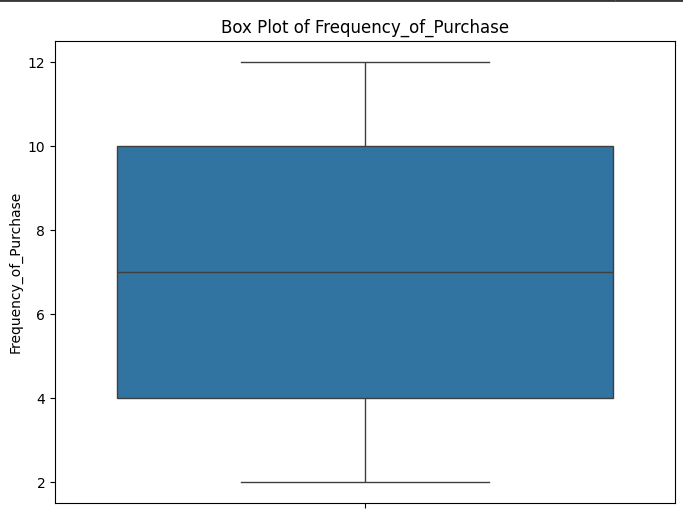


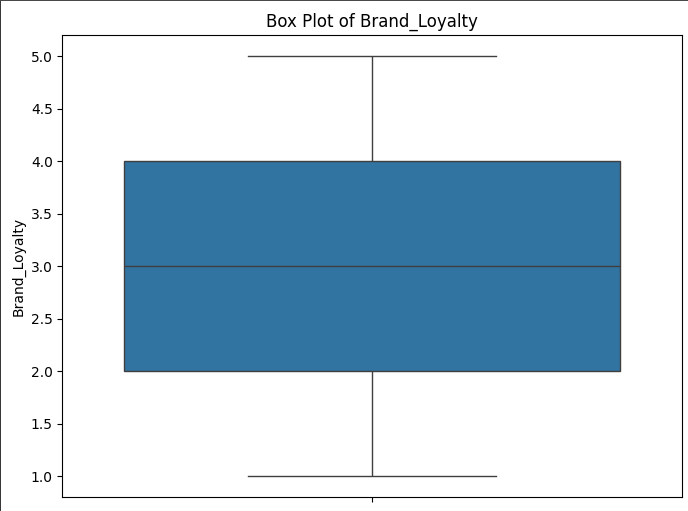


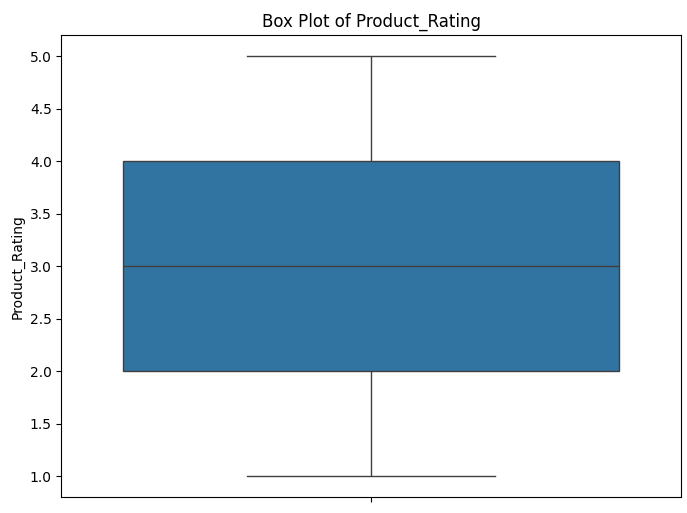


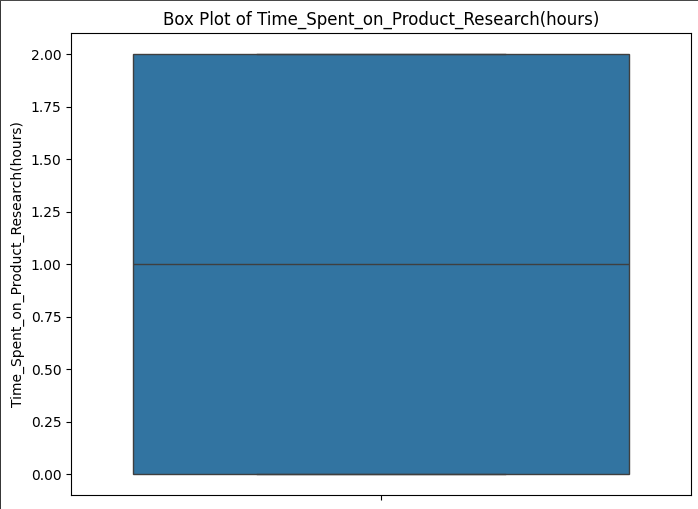


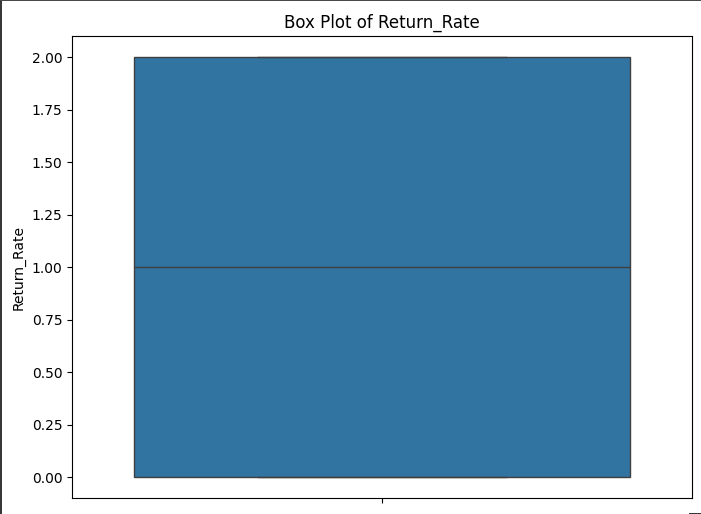


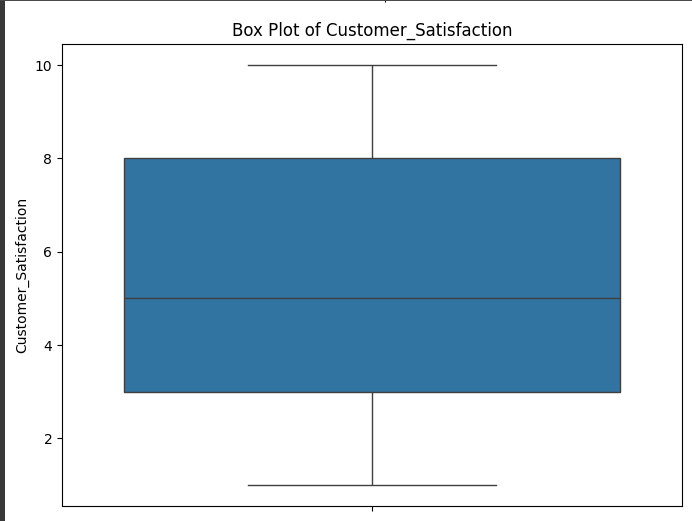


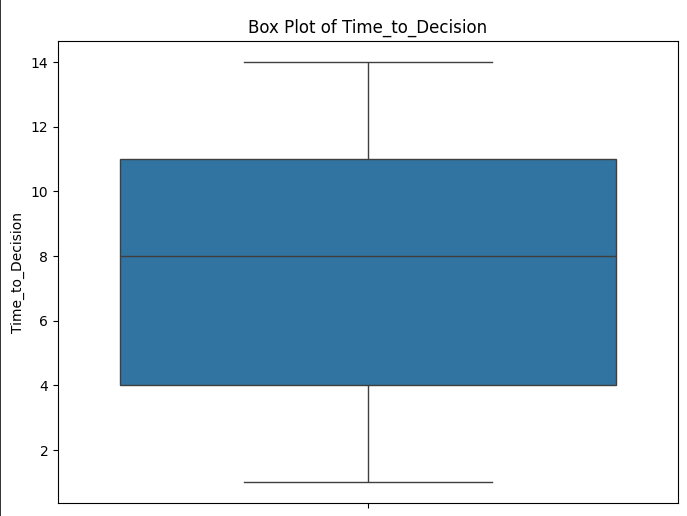


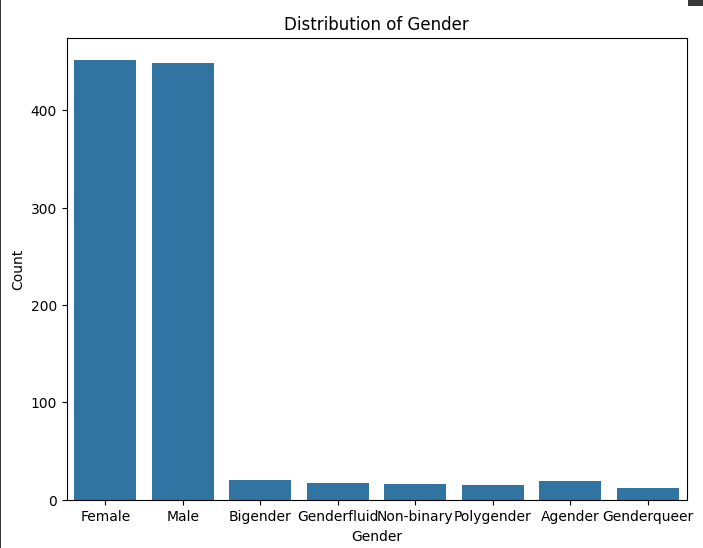






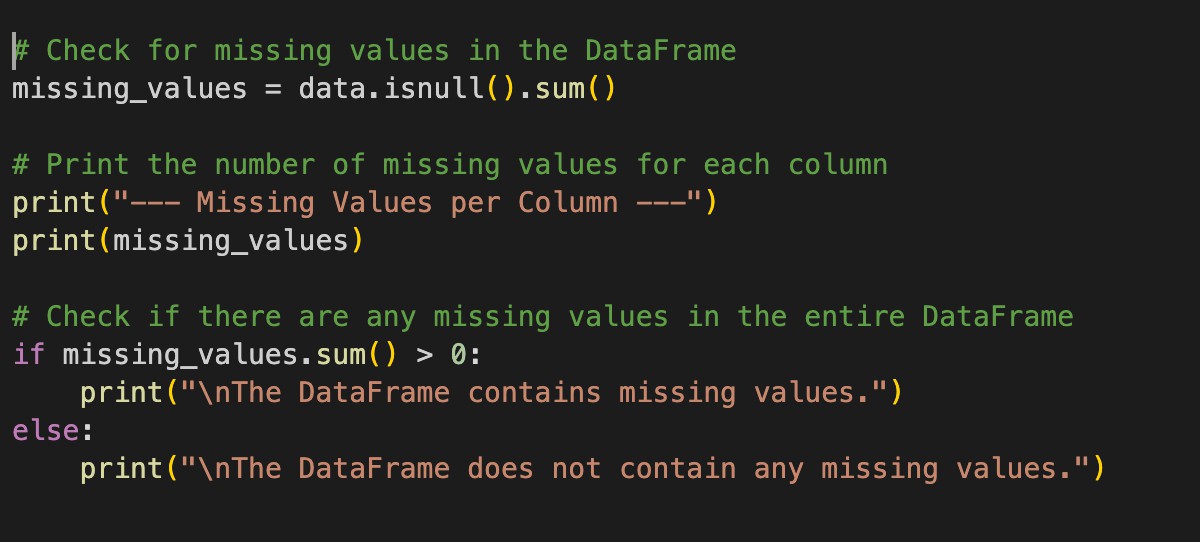




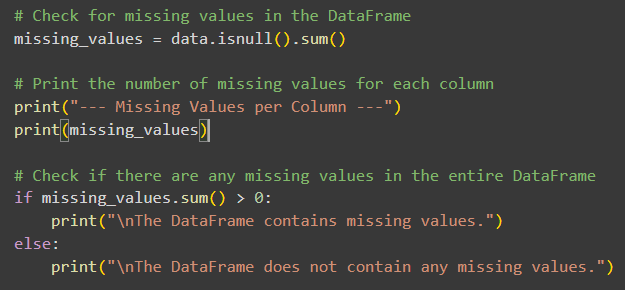


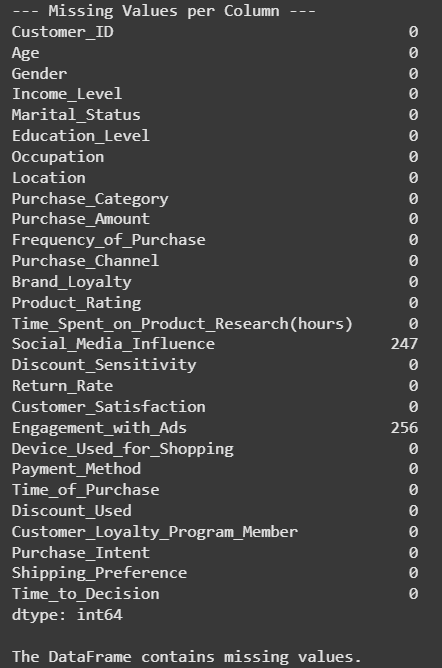
Kode pada bagian ini berfungsi untuk menyajikan data dalam bentuk visual agar lebih mudah dipahami. Pertama, data dibaca dari Google Drive menggunakan library pandas, kemudian divisualisasikan dengan matplotlib dan seaborn. Histogram digunakan untuk melihat distribusi data numerik beserta kepadatan (KDE), sedangkan box plot menampilkan sebaran data dan mendeteksi adanya outlier. Scatter plot memperlihatkan hubungan antara dua variabel, contohnya antara Age dan Purchase Amount (USD), sehingga dapat diketahui apakah ada pola tertentu di antara keduanya. Sementara itu, bar plot digunakan untuk menunjukkan jumlah data berdasarkan kategori, seperti distribusi jenis kelamin pada kolom Gender.

# Bagian 6: Mengecek Data Kosong

Pada bagian ini akan dicontohkan bagaimana cara mengecek apakah ada data kosong atau tidak.

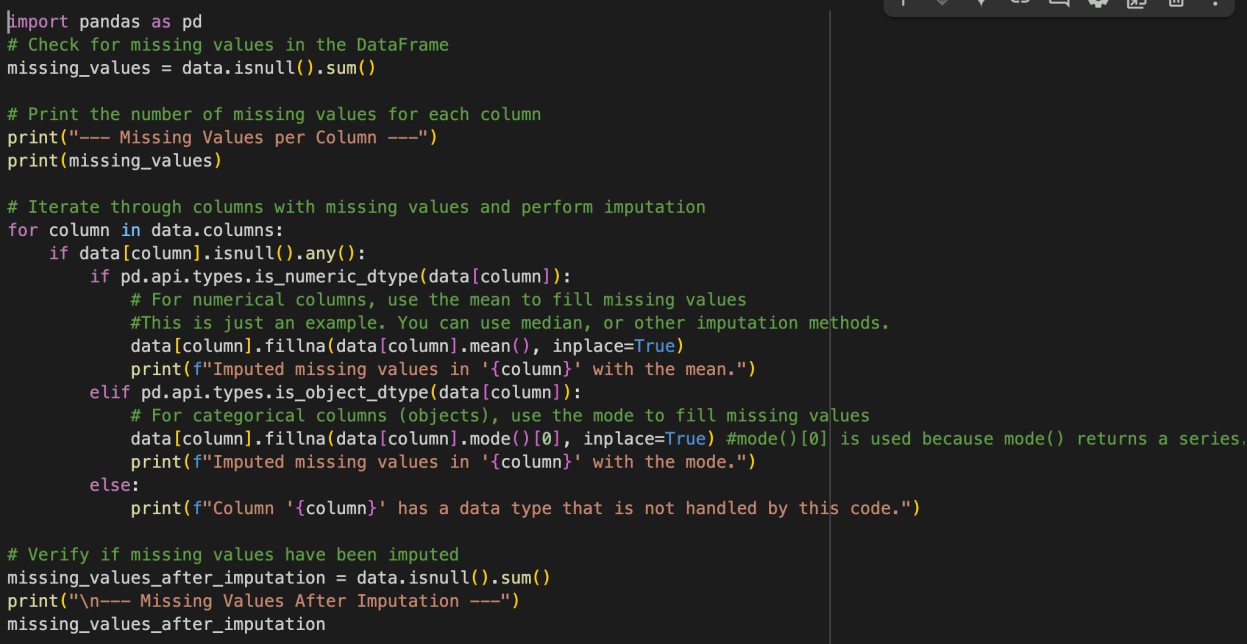
Silakan dicoba, bagaimana hasilnya?



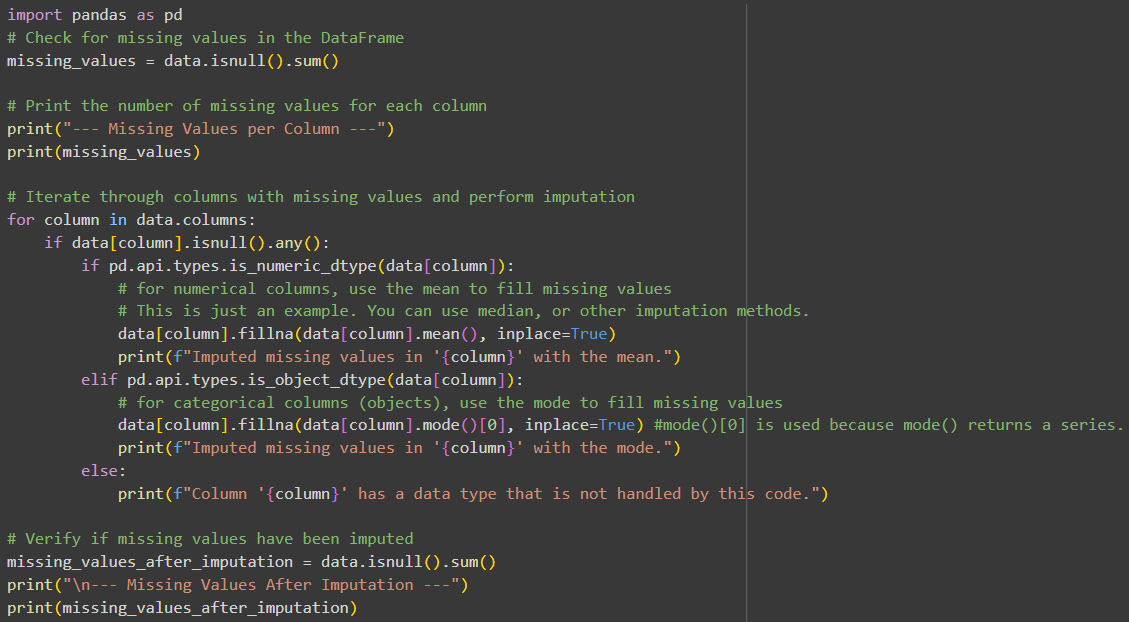


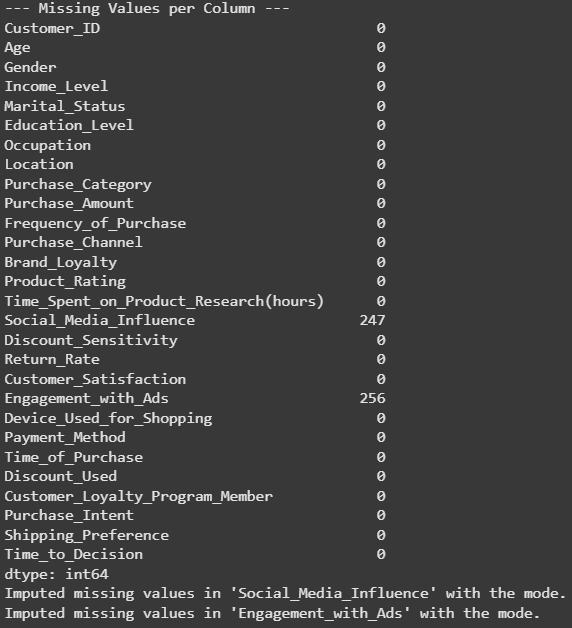
Perintah data.isnull().sum() akan menghitung jumlah nilai kosong pada setiap kolom dalam DataFrame, kemudian hasilnya ditampilkan agar dapat diketahui kolom mana saja yang memiliki data hilang. Jika total nilai kosong lebih dari nol, program menampilkan pesan “The DataFrame contains missing values”, menandakan bahwa ada kolom yang belum terisi sepenuhnya. Sebaliknya, jika tidak ada nilai kosong, program menampilkan pesan “The DataFrame does not contain any missing values”. Dari output ini, pengguna dapat dengan mudah mengidentifikasi bagian data yang perlu dilakukan *imputasi* pada tahap berikutnya.

# Bagian 7: Treatment Data Kosong

Jika ada data kosong, maka dilakukan imputasi data, berikut ini cara melakukannya.

Silakan dicoba, bagaimana hasilnya?





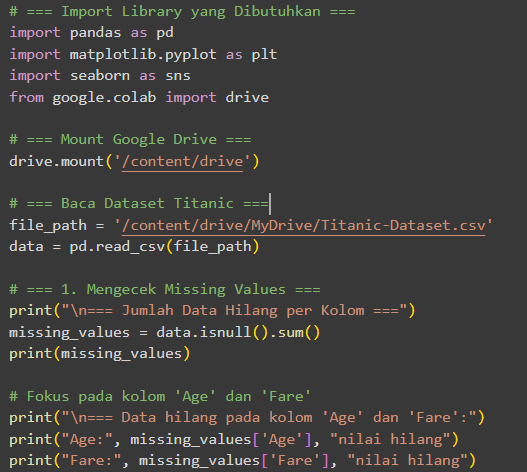


Kode di atas digunakan untuk melakukan imputasi data kosong (missing values) pada dataset, yaitu proses mengganti nilai yang hilang dengan nilai yang sesuai agar data dapat diolah tanpa error. Langkah pertama, program memeriksa setiap kolom menggunakan data.isnull().sum() untuk mengetahui jumlah nilai kosong. Kemudian, untuk kolom numerik, nilai kosong diisi menggunakan mean (rata-rata) dari kolom tersebut, sedangkan untuk kolom kategorikal (tipe object), nilai kosong diisi dengan mode (nilai yang paling sering muncul). Setelah proses imputasi dilakukan, kode kembali memeriksa apakah masih ada data kosong menggunakan data.isnull().sum(). Hasil output menunjukkan bahwa semua kolom yang awalnya memiliki nilai hilang kini telah terisi, ditandai dengan nilai 0 pada jumlah *missing values* setelah imputasi. Dengan demikian, dataset sudah bersih dan siap digunakan untuk analisis statistik selanjutnya.

# Tugas

## Tugas 1

* Download dataset Titanic pada [https://www.kaggle.com/datasets/yasserh/titanic-](https://www.kaggle.com/datasets/yasserh/titanic-dataset/code) [dataset/code](https://www.kaggle.com/datasets/yasserh/titanic-dataset/code)
* Cek, apakah variabel “Age” dan “Fare” memiliki data yang nilainya hilang. Jika ya, berapa jumlahnya?

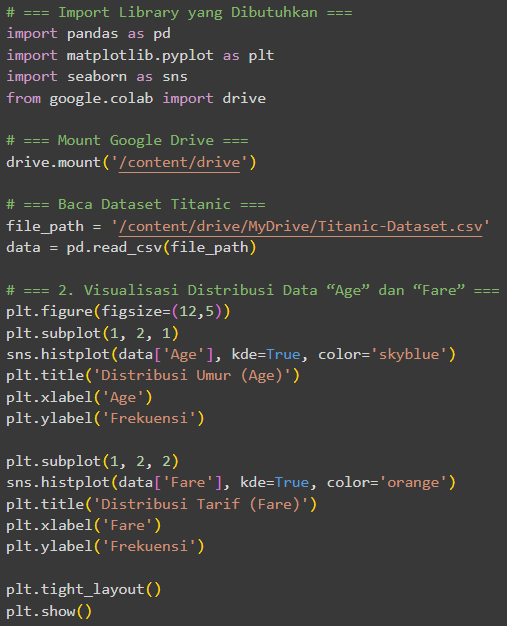


**Output**

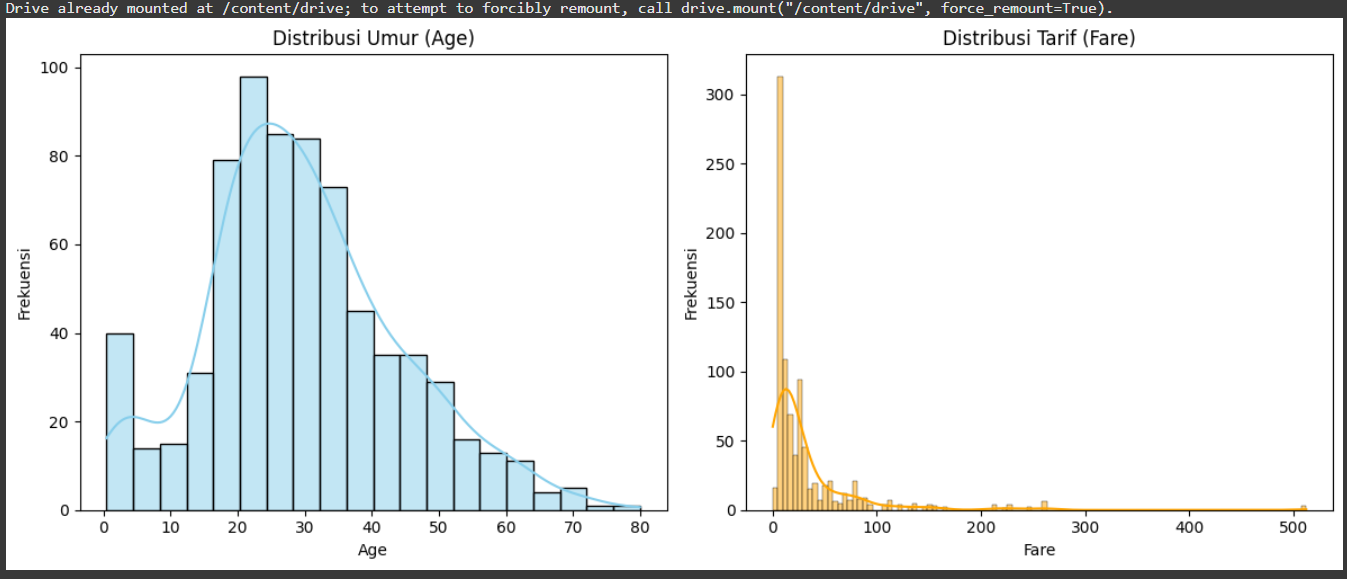
****

Berdasarkan hasil pemeriksaan menggunakan fungsi data.isnull().sum(), ditemukan bahwa kolom Age memiliki sejumlah nilai yang hilang, sedangkan kolom Fare sebagian besar terisi lengkap.

* Cek pola distribusi data pada “Age” dan “Fare”.

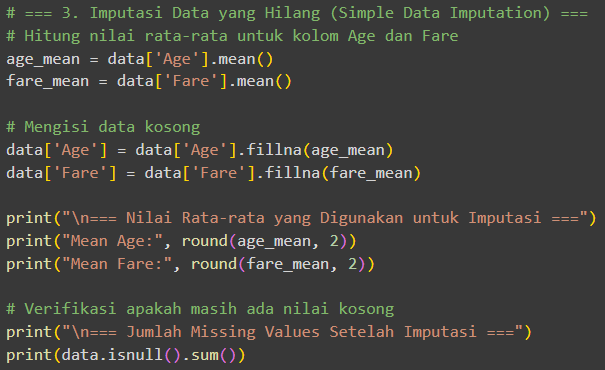


**Output**

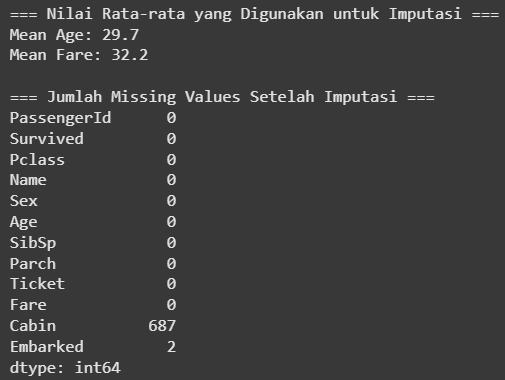
****

Untuk variabel Age, distribusi data terlihat mendekati normal dengan puncak di sekitar usia 20–35 tahun, yang menandakan sebagian besar penumpang Titanic berusia muda hingga dewasa awal. Namun terdapat sedikit ekor panjang ke arah kanan (positif skew), mengindikasikan adanya beberapa penumpang dengan usia yang jauh lebih tua dari rata-rata. Sementara itu, distribusi tarif (Fare) menunjukkan pola yang sangat miring ke kanan (right-skewed). Sebagian besar penumpang membayar tiket dengan harga rendah, yaitu di bawah 100 unit tarif, dan hanya sebagian kecil penumpang yang membayar harga tiket sangat tinggi hingga lebih dari 500 unit tarif. Pola ini mencerminkan adanya perbedaan kelas sosial di kapal Titanic, di mana mayoritas penumpang berasal dari kelas ekonomi menengah ke bawah.

* Gunakan teknik *simple data imputation* yang tepat untuk mensubtitusi data yang hilang.



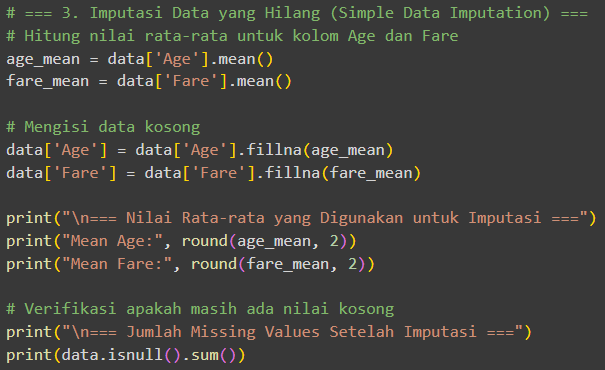
**Output**

****

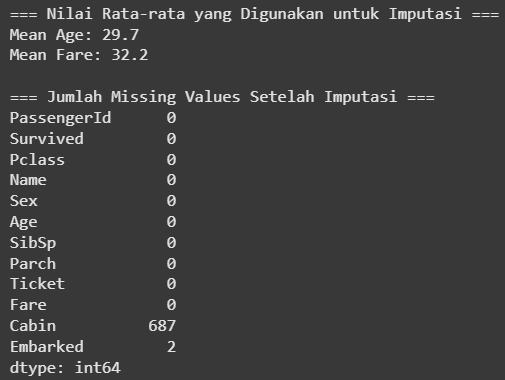
Pada kolom Age dan Fare dalam dataset Titanic. Pertama, menghitung nilai rata-rata dari kedua kolom tersebut menggunakan fungsi mean(), kemudian nilai kosong diisi dengan rata-rata yang sesuai menggunakan fillna(). Setelah proses imputasi, program menampilkan nilai rata-rata yang digunakan sebagai pengganti serta memeriksa kembali jumlah data kosong di setiap kolom menggunakan data.isnull().sum(). Hasil output menunjukkan bahwa nilai kosong pada Age dan Fare telah berhasil digantikan oleh nilai rata-rata, misalnya Mean Age: 29.7 dan Mean Fare: 32.2 sehingga tidak ada lagi data hilang pada dataset. Langkah ini memastikan data menjadi lengkap tanpa gangguan missing values.

## Tugas 2

* Imputasi data yang hilang pada “Age” dengan menggunakan mean

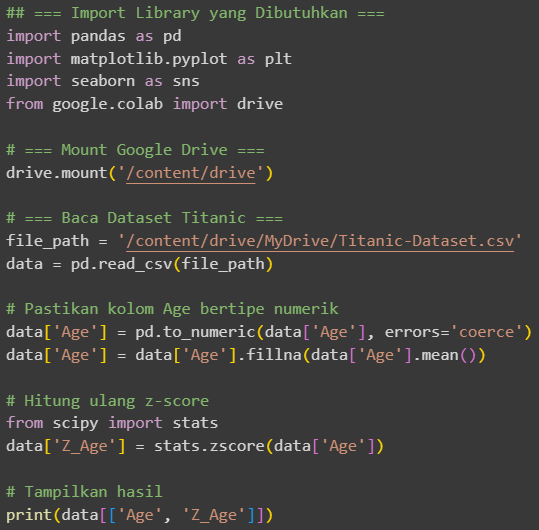


**Output**

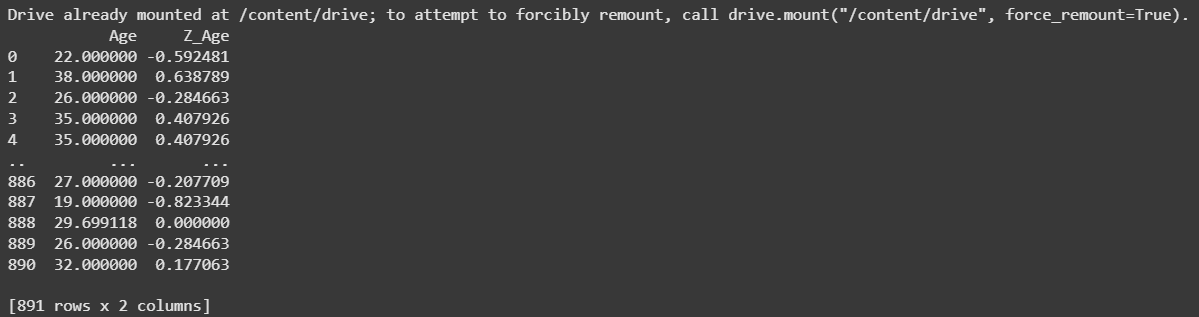
****

Pada tugas ini dilakukan analisis terhadap variabel Age pada dataset Titanic. Nilai kosong pada kolom Age diimputasi menggunakan rata-rata (mean) sebesar ≈ 29.7 tahun agar tidak ada lagi data hilang. Pada soal sebelumnya saya sudah mengisi missing values termasuk Age dengan menggunakan mean nya, oleh karena itu langsung lanjut ke nomor berikutnya untuk menghitung z-score dari “Age” untuk semua data.

* Hitung z-score dari “Age” untuk semua data



**Output**

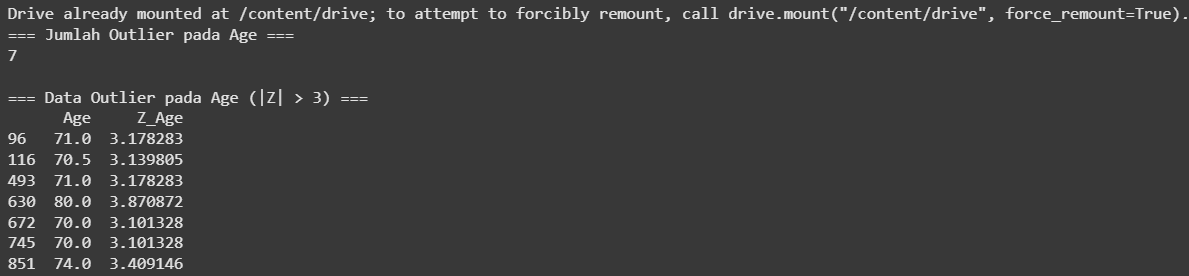
****

Hasil perhitungan Z-Score pada kolom Age menunjukkan sejauh mana setiap nilai umur menyimpang dari rata-rata dalam satuan deviasi standar. Nilai Z-Score positif menandakan usia di atas rata-rata (29.7 tahun), sedangkan nilai negatif menunjukkan usia di bawah rata-rata. Contohnya, umur 22 tahun memiliki Z-Score –0.59 yang berarti lebih muda dari rata-rata dan umur 38 tahun memiliki Z-Score 0.63 yang berarti lebih tua dari rata-rata. Z-Score mempermudah identifikasi posisi setiap nilai umur terhadap rata-rata dan menjadi dasar untuk mendeteksi outlier atau usia ekstrem di step selanjutnya.

* Tentukan jumlah outlier dari “Age”  |𝑍| > 3

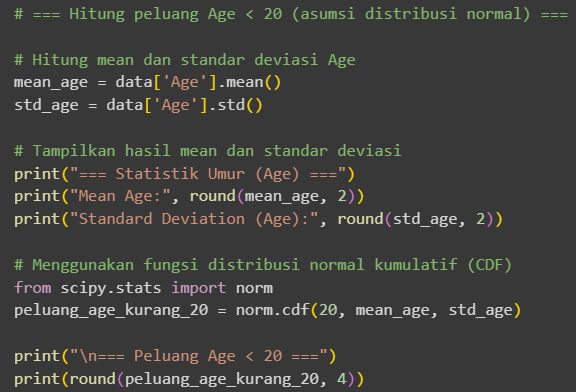
****

**Output**

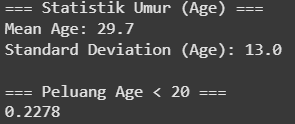
****

Selanjutnya dilakukan perhitungan Z-Score untuk setiap data Age menggunakan rumus , sehingga diperoleh kolom baru Z\_Age yang menunjukkan jarak setiap nilai umur terhadap rata-rata dalam satuan deviasi standar. Berdasarkan hasil perhitungan, terdapat 7 data dengan nilai |Z| > 3, yang dikategorikan sebagai outlier, menunjukkan adanya penumpang dengan usia jauh lebih tinggi dibandingkan mayoritas yang berada pada rentang 70–80 tahun.

* Diasumsuikan ”Age” terdistribusi secara normal, berapa peluang "𝐴𝑔𝑒” < 20?



**Output**

****

Dengan asumsi distribusi normal berparameter mean = 29.7 dan standar deviasi = 13.0, diperoleh peluang atau sekitar 22.78%, yang berarti sekitar seperlima penumpang diperkirakan berusia di bawah 20 tahun.