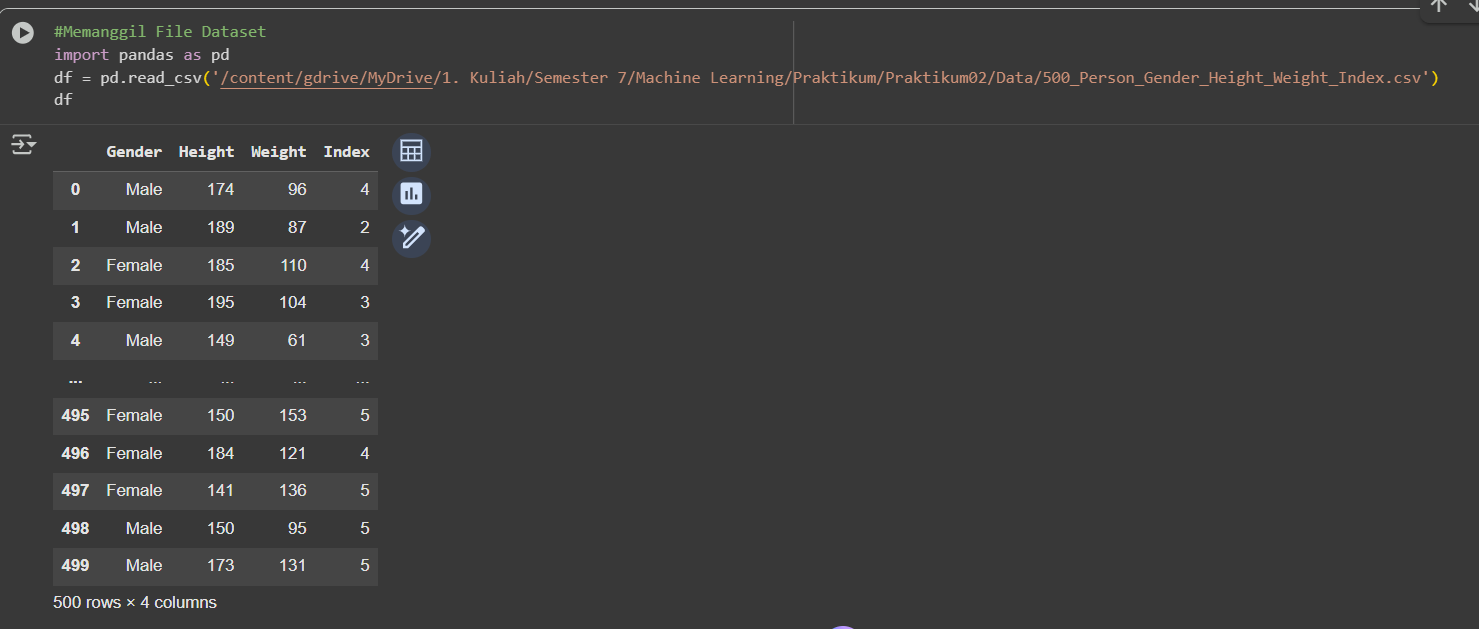
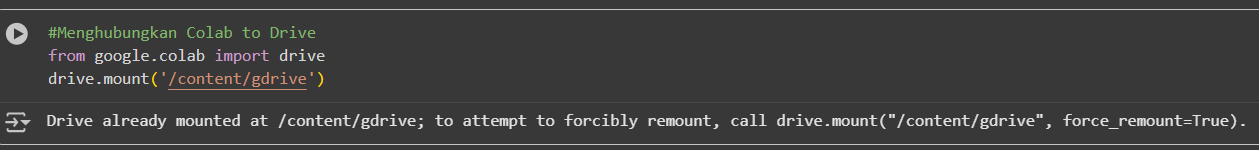
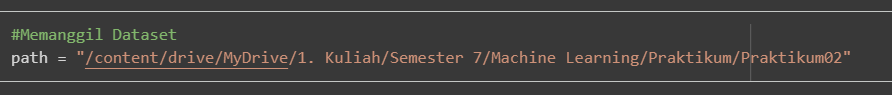
**Praktikum 2 – Machine Learning**

**Rahmi Atika - 0110222279**

Teknik Informatika, STT Terpadu Nurul Fikri, Depok

E-mail:  [rahm22279ti@student.nurulfikri.ac.id](mailto:ahm22279ti@student.nurulfikri.ac.id)

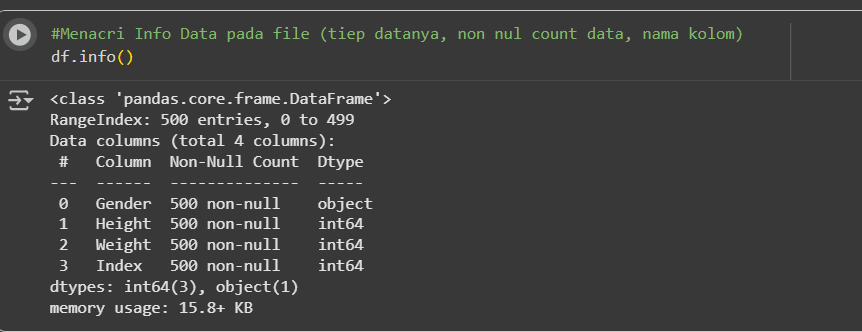
1. **Hasil Praktikum Mandiri**



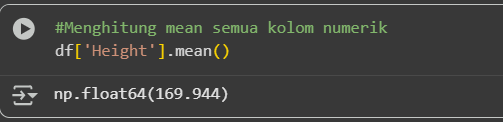
Kode tersebut digunakan untuk menghubungkan Google Colab dengan Google Drive. `Setelah kode dieksekusi, akan muncul tautan otorisasi untuk memberikan izin agar Google Colab dapat mengakses file yang tersimpan di Google Drive.

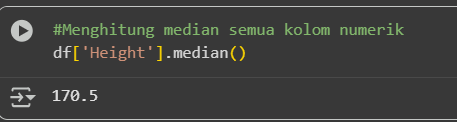
Kode Import pandas as pd digunakan untuk mengimpor Library pandas dengan alias pd.

Kode df = pd.read\_csv digunakan untuk membaca file CSV yang berada di folder data, setelah itu disimpan sebagai df (DateFrame). Variabel path berisi lokasi folder di Google Drive tempat file 500\_Person\_Gender\_Height\_Weight\_Index.csv disimpan.

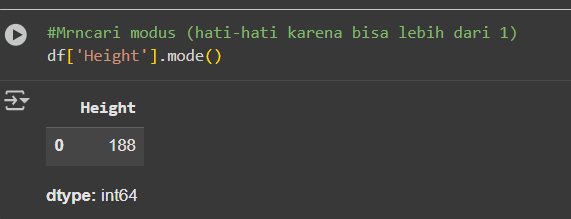


Kode [df.info](http://df.info)() digunakan untuk menampilkan ringkasan singkat mengenai isi data dalam dataframe seperti Gender dengan tipe data object, Height dengan tipe data Int64 dan yang lainnya.

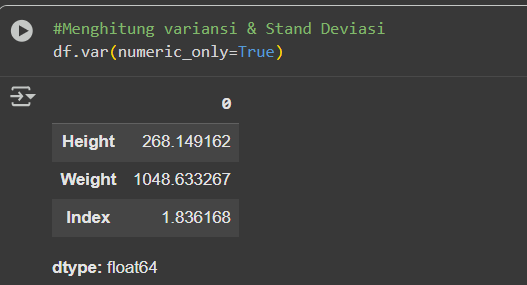
Kode .mean() digunakan untuk menghitung rata-rata (jumlah semua nilai dibagi banyaknya data) untuk setiap kolom numerik dalam DataFrame.



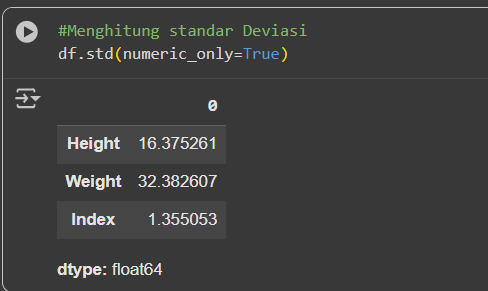
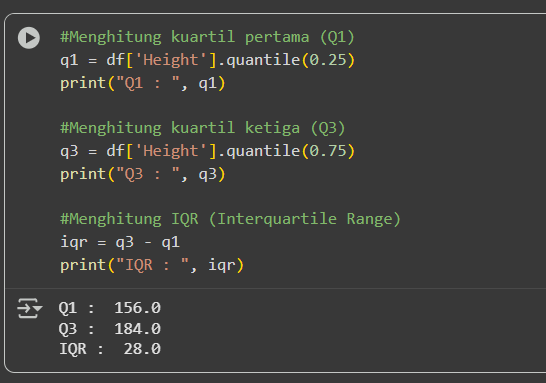
Kode .median() digunakan untuk menghitung nilai tengah dari data yang sudah diurutkan. Jika jumlah data genap maka nilai mediannya yaitu rata-rata dari dua nilai tengah.



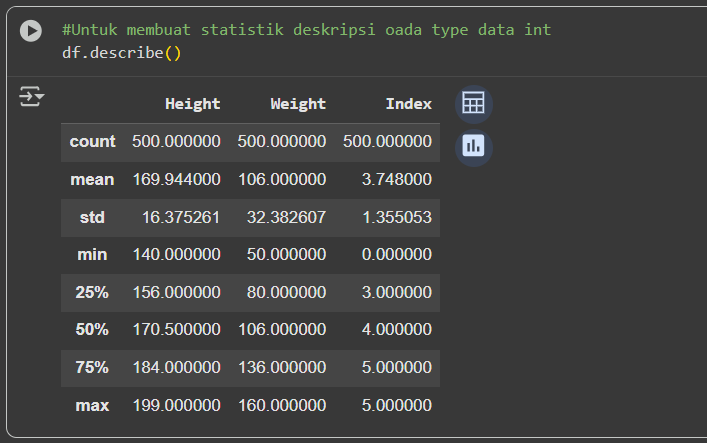
Kode .mode().iloc[0] digunakan untuk mengambil modus atau nilai yang paling sering muncul. Karena .mode() bisa menghasilkan lebih dari satu nilai (jika ada lebih dari satu modus), maka .iloc[0] digunakan untuk mengambil modus pertama.

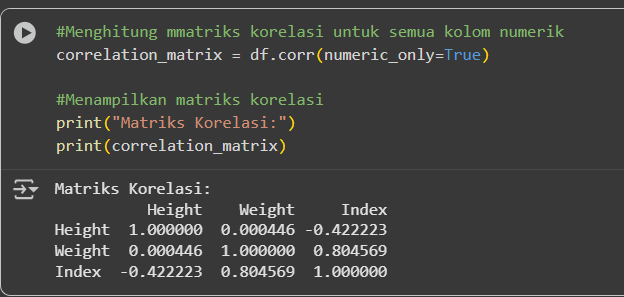


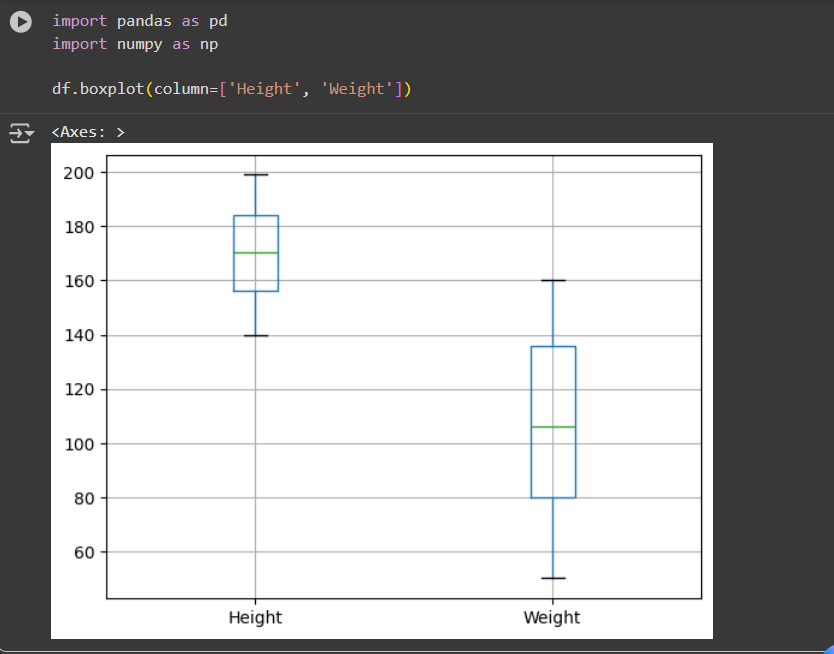
Kode .var() atau Variansi digunakan untuk menghitung atau melihat ukuran persebaran seberapa jauh nilai-nilai data menyebar dari rata-ratanya.

Kode .std() atau Standar Deviasi digunakan untuk mengakar-kuadratkan variansi, yang memberikan ukuran persebaran dalam satuan yang sama dengan data aslinya. 

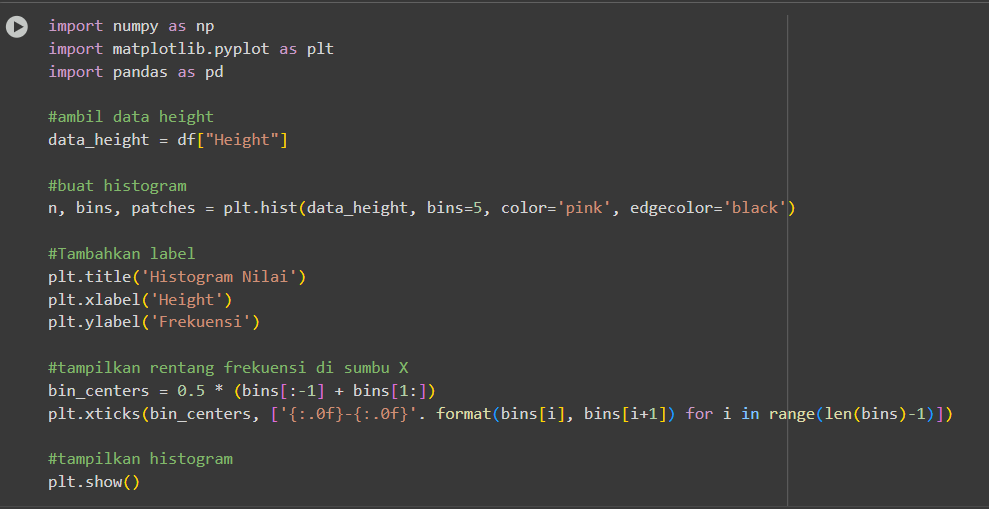
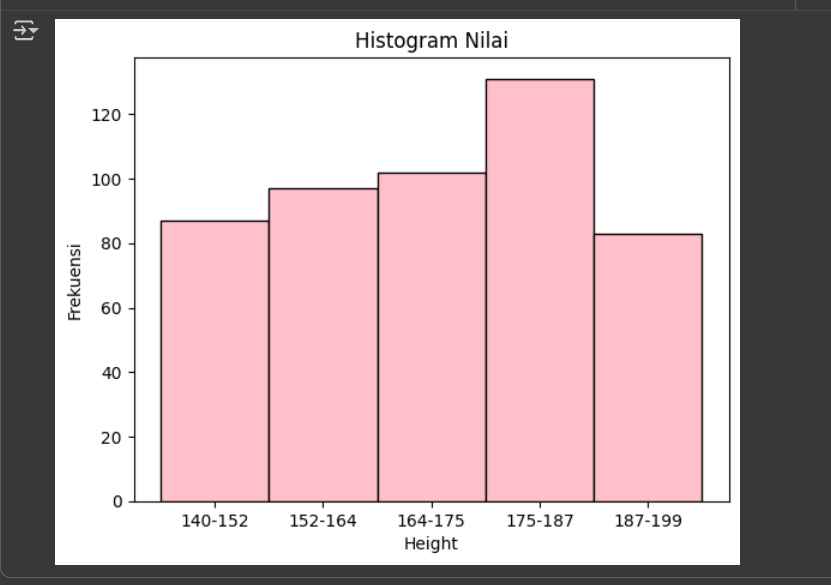
Kode df['Height'].quantile(0.25) dan df['Height'].quantile(0.75) digunakan untuk menghitung kuartil dan IQR dari data Height. Kuartil pertama (Q1) adalah nilai pada 25% data terbawah, sedangkan kuartil ketiga (Q3) pada 75% data teratas. IQR diperoleh dari selisih Q3 − Q1 dan menunjukkan lebar sebaran data di bagian tengah.

Kode .describe() digunakan untuk menghitung statistik deskriptif dasar dari setiap kolom numerik pada DataFrame. Hasil yang ditampilkan berupa tabel yang berisi beberapa metrik, seperti jumlah data non-null (count), nilai rata-rata (mean), standar deviasi (std), nilai terkecil (min), kuartil pertama (25%), median atau kuartil kedua (50%), kuartil ketiga (75%), serta nilai terbesar (max).

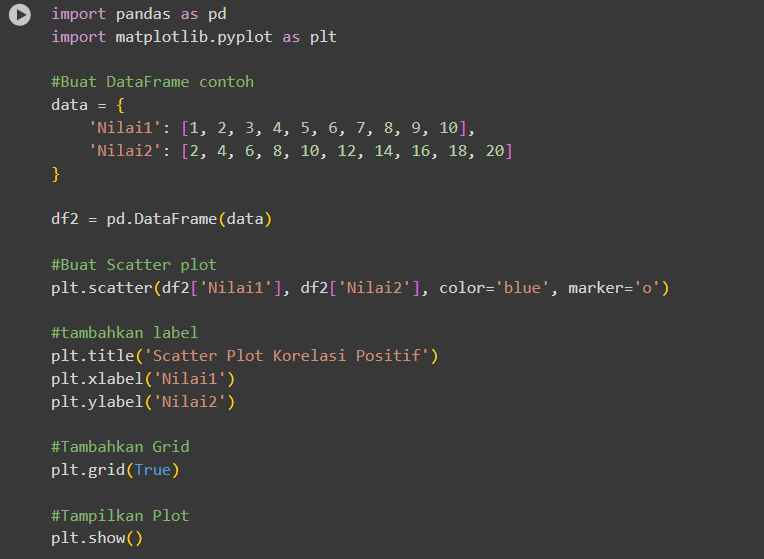
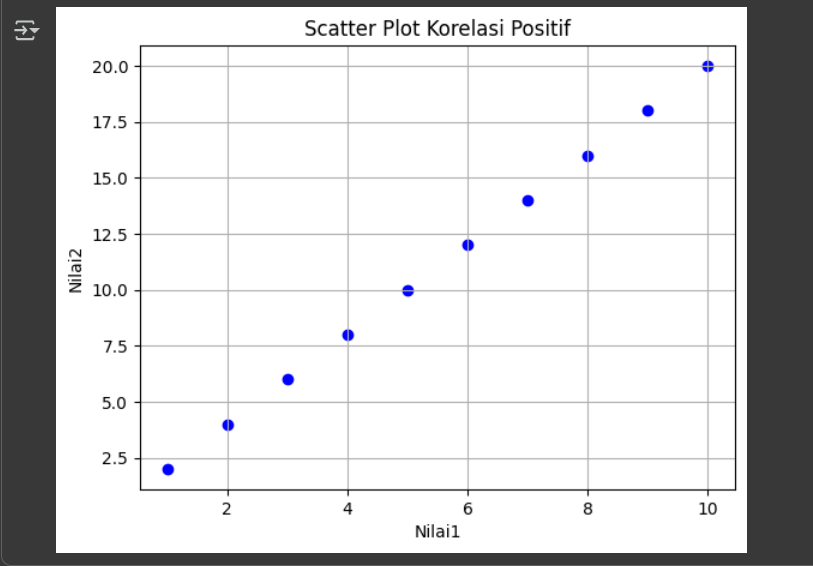
Kode tersebut digunakan untuk menghitung koefisien korelasi pearson semua kolom numerik dalam DataFrame atau df untuk melihat seberapa kuat hubungan linear antara 2 variabel tersebut dengan nilai korelasi -1 hingga 1. Jika nilai korelasi mendekasi 1 maka menunjukkan korelasi positif yaitu di mana satu variabel naik jika variabel lain juga naik. Sedangkan jika nilai korelasi mendekati -1 maka menunjukkan korelasi negatif yaitu di mana satu variabel naik namun variabel lain turun. Dan jika nilai korelasi mendekati 0 maka itu menunjukkan tidak adanya hubungan linear.

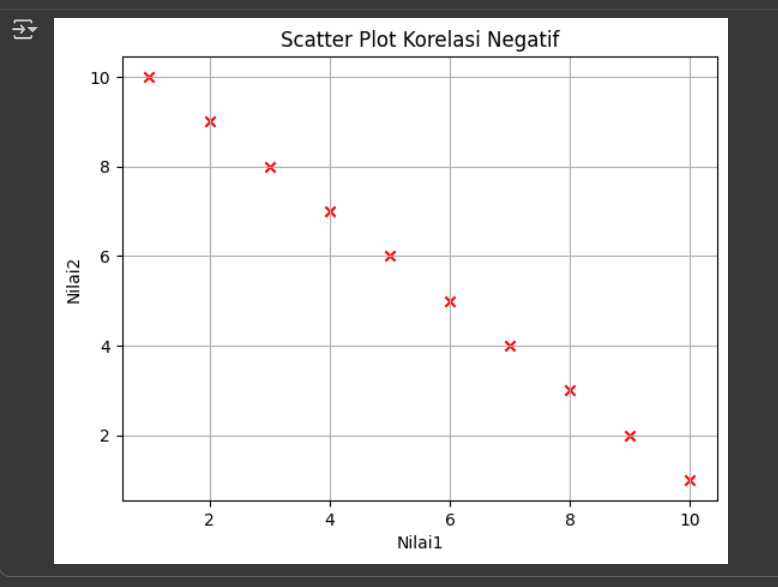
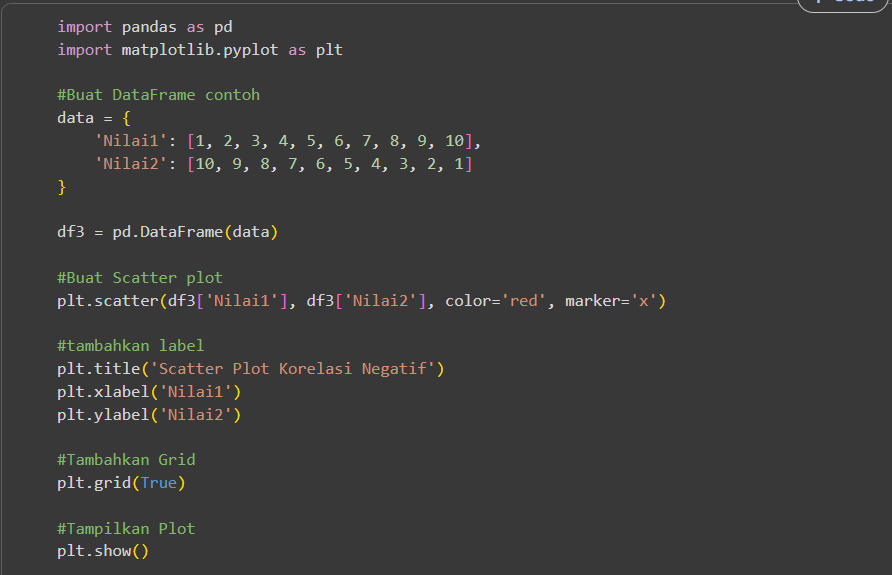


Kode tersebut menggunakan fungsi df.boxplot() untuk membuat visualisasi data dalam bentuk boxplot pada kolom Height dan Weight. Boxplot berfungsi untuk memberikan gambaran ringkas mengenai distribusi data. Garis di tengah kotak menggambarkan median atau nilai tengah, sedangkan ukuran kotak merepresentasikan Interquartile Range (IQR), yaitu rentang antara kuartil pertama dan kuartil ketiga. Garis horizontal di bagian bawah dan atas kotak, yang disebut whiskers, menunjukkan batas nilai yang dianggap masih wajar dalam data.



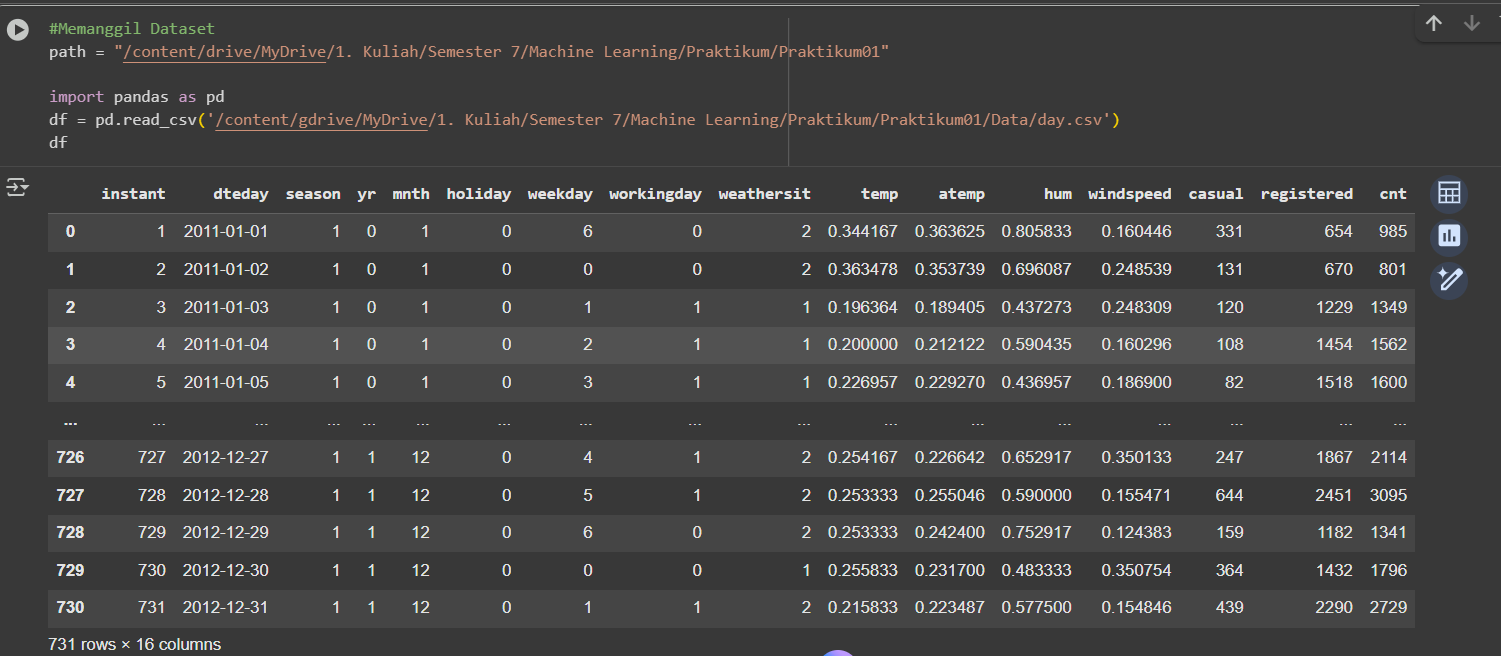
Histogram pada kolom Height menampilkan sebaran tinggi badan dalam 5 rentang nilai (bins). Setiap batang menunjukkan jumlah data pada rentang tertentu, misalnya rentang 175–187 memiliki frekuensi tertinggi.

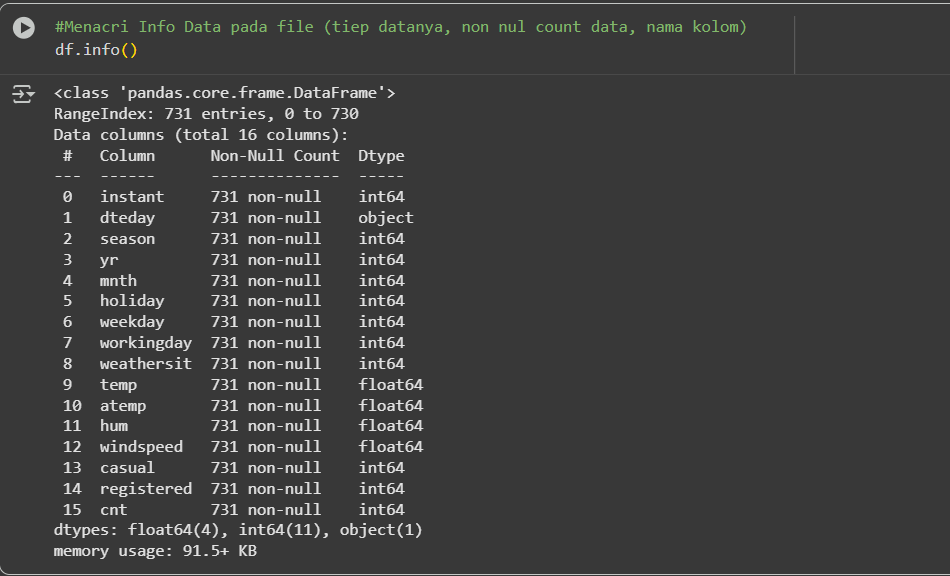
Scatter plot digunakan untuk menggambarkan hubungan antara dua variabel numerik melalui titik-titik pada grafik. Pada contoh, variabel Nilai1 dan Nilai2 membentuk pola titik yang naik dari kiri bawah ke kanan atas, menunjukkan adanya korelasi positif, yaitu ketika Nilai1 meningkat maka Nilai2 juga ikut meningkat

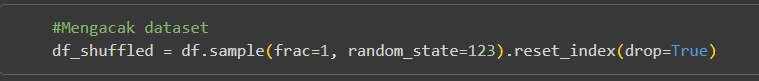


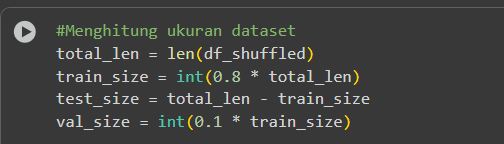
Sedangkan pada hasil yang kedua ini Scatter Plot menunjukkan Korelasi Negatif, yaitu ketika titik-titik pada grafik menunjukkan Nilai1 bertambah, Nilai2 justru menurun, membentuk pola menurun dari kiri atas ke kanan bawah. Hal ini menandakan adanya hubungan terbalik antara kedua variabel.

Tugas Mandiri

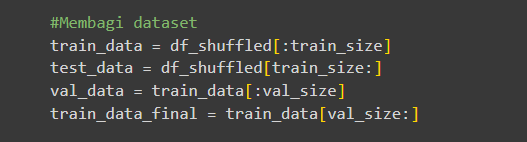
Variabel path digunakan untuk menyimpan lokasi direktori tempat dataset disimpan dalam drive. Kode import pandas as pd digunakan untuk mengimpor library pandas dengann alias pd. Pandas merupakan sebuah library pada Python dalam pengolahan data tabular untuk format file seperti CSV yang berguna untuk membaca, membersihkan, dan menganalisis data.

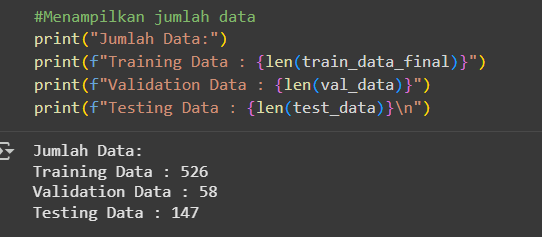
Kode df.info() digunakan untuk melihat ringkasan struktur dataset. Dari hasilnya terlihat bahwa dataset memiliki 731 baris dan 16 kolom. Semua kolom berisi 731 non-null, artinya tidak ada data yang kosong. Tipe data terdiri dari int64 (angka bulat), float64 (angka desimal), dan object (teks, misalnya tanggal pada kolom dteday)

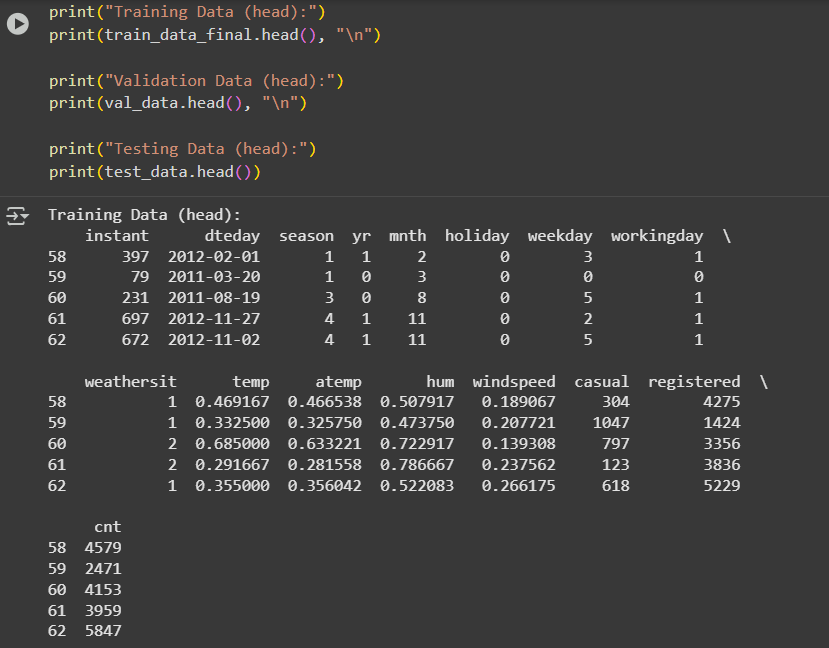
Dataset perlu diacak agar data yang akan dipisah menjadi data training, validation dan testing tidak dipengaruhi oleh urutan data aslinya. Fungsi sample(frac=1) digunakan untuk mengacak seluruh isi data. Parameter random\_state=123 digunakan agar hasil acakan selalu sama setiap kali kode dijalankan. Setelah data diacak, indeks barisnya diatur ulang dengan reset\_index(drop=True) agar kembali rapi dimulai dari 0.

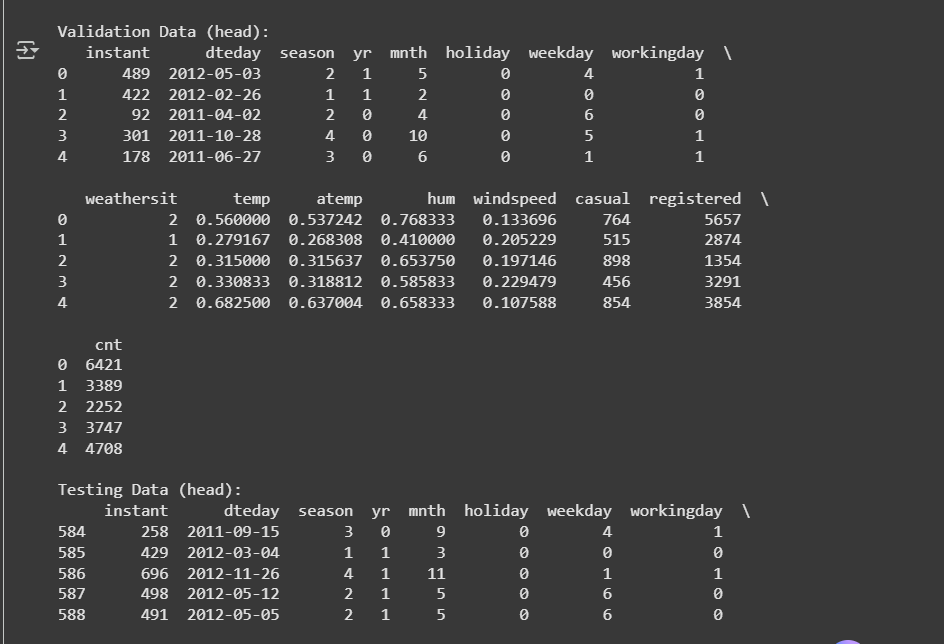
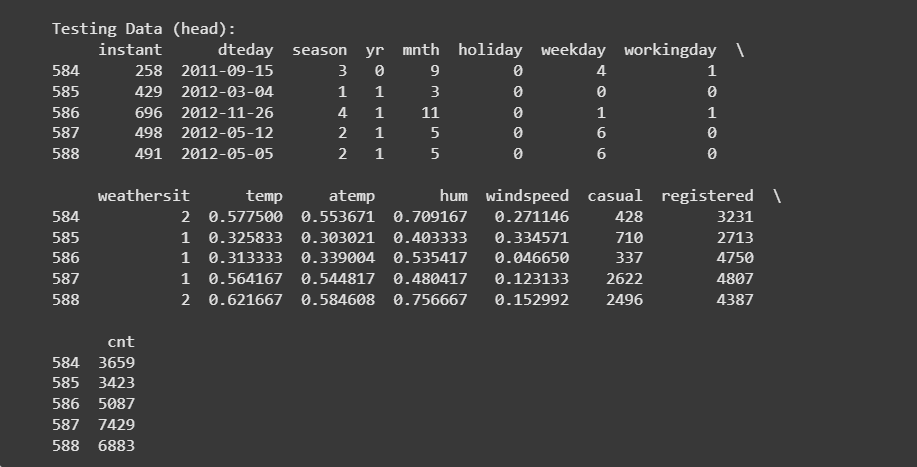


Kode ini dipakai untuk menghitung berapa banyak data yang akan masuk ke tiap bagian. Variabel total\_len menyimpan jumlah seluruh baris data. Dari total tersebut, 80% disiapkan untuk data training (train\_size), sementara sisanya 20% dijadikan data testing (test\_size). Lalu, dari data training itu sendiri, diambil lagi 10% (val\_size) untuk dipakai sebagai data validation.

Kode tersebut digunakan untuk memisahkan data menjadi beberapa bagian. Sebanyak 80% data pertama dimasukkan ke dalam train\_data, sedangkan 20% sisanya digunakan sebagai test\_data. Dari train\_data itu sendiri, 10% pertama dipisahkan menjadi val\_data sebagai data validasi, sementara sisanya menjadi train\_data\_final, yaitu data training utama yang siap digunakan.

Kode tersebut menampilkan jumlah baris dari tiap bagian data setelah dibagi. Hasilnya, data training berisi 526 baris, validation 58 baris, dan testing 147 baris

Pada 2 kode print pertama menampilkan 5 baris pertama dari data training. Terlihat ada kolom seperti tanggal, musim, cuaca, suhu, kelembapan, dan jumlah pengguna sepeda. Data ini adalah 80% dari dataset (dikurangi 10% untuk validasi) yang nantinya dipakai untuk melatih model.

Pada 2 kode print berikutnya menampilkan 5 baris pertama dari data validation. Data ini diambil dari 10% data training yang sudah dipisahkan tadi. Fungsi dari kode tersebut yaitu untuk mengecek apakah model yang dilatih bisa bekerja baik pada data yang tidak dipakai langsung saat training.

Lalu pada 2 kode terakhir terakhir memperlihatkan 5 baris pertama dari data testing. Dataset ini berjumlah 20% dari total data dan hanya digunakan pada tahap akhir untuk menilai kinerja model secara objektif.

**Link GitHub**

Praktikum:

<https://github.com/Rhmiatika/TI-14_Rahmi_0110222279/blob/a57910be1ff1aab918965fcb4c66fa35232d90f2/Praktikum01/Notebook/praktikum01.ipynb>

Tugas Mandiri:

<https://github.com/Rhmiatika/TI-14_Rahmi_0110222279/blob/a57910be1ff1aab918965fcb4c66fa35232d90f2/Praktikum01/Notebook/tugasmandiri1.ipynb>