Korelasi dan Visualisasi Lanjutan

dalam EDA

Setelah memahami dasar-dasar EDA dan statistika deskriptif, mari kita perdalam pengetahuan dengan konsep korelasi dan beberapa teknik visualisasi tambahan yang berguna.

# Korelasi

Korelasi adalah ukuran statistika yang menunjukkan kekuatan dan arah hubungan linear antara dua variabel. Nilai korelasi berkisar antara -1 hingga 1.

* Korelasi Positif (0 < r ≤ 1): Kedua variabel cenderung meningkatkan atau menurun bersama-sama.
* Korelasi Negatif (-1 ≤ r < 0): Kedua variabel cenderung bergerak berlawanan arah (satu meningkat, yang lain menurun).
* Tidak Ada Korelasi (r = 0): Tidak ada hubungan linear antara kedua variabel.

# Mengukur Korelasi

Salah satu cara paling umum untuk mengukur korelasi adalah menggunakan koefisien korelasi Pearson(r). Pandas menyediakan fungsi .corr() untuk menghitung korelasi antara kolom-kolom numerik dalam DataFrame.

# Visualisasi Korelasi

Salah satu cara memvisualisasi korelasi adalah menggunakan **heatmap.** Heatmap adalah representasi grafis dari matriks korelasi, dimana nilai korelasi direpresentasikan dengan warna. Semakin gelap warnanya, semakin kuat korelasinya.

# Visualisasi Tambahan

Selain heatmap, ada beberapa visualisasi lain yang berguna dalam EDA:

* **Pair Plot:** Membuat scatter plot antara setiap pasangan kolom numerik dalam DataFrame.
* **Bar Plot:** Membandingkan nilai rata-rata atau jumlah dari variabel kategorikal.
* **Pie Chart:** Menampilkan proporsi dari setiap kategori dalam variabel kategorikal.
* **Line Plot:** Menampilkan perubahan suatu variabel terhadap waktu.

## Tugas Hari 5:

Lanjutkan dengan dataset Titanic.

**Tugas:**

1. **Hitung Korelasi:** Gunakan untuk menghitung korelasi antara semua pasangan kolom numerik dalam dataset Titanic. Tampilkan hasilnya dalam bentuk tabel.
2. **Visualisasikan Korelasi:** Buat heatmap untuk memvisualisasikan matriks korelasi yang telah Anda hitung.
3. **Pair Plot:** Buat pair plot untuk semua kolom numerik dalam dataset Titanic.
4. **Analisis:** Berikan interpretasi singkat tentang hasil korelasi dan visualisasi yang Anda buat. Apakah ada pasangan variabel yang memiliki korelasi kuat? Apakah ada hubungan yang menarik yang Anda temukan?

### **Contoh Kode (pandas dan seaborn):**

import pandas as pd

import seaborn as sns

import matplotlib.pyplot as plt

# ... (kode untuk membaca dataset Titanic)

# Hitung korelasi

correlation\_matrix = titanic\_df.corr()

print(correlation\_matrix)

# Heatmap

sns.heatmap(correlation\_matrix, annot=True)

plt.title('Korelasi Antar Variabel')

plt.show()

# Pair plot

sns.pairplot(titanic\_df)

plt.show()

**Catatan:** Jangan lupa untuk mengimport library seaborn untuk membuat heatmap dan pair plot.

**Selamat mengerjakan tugas!!! 🙂**

Tugas:

1. Hitung Korelasi

# Hitung Korelasi

correlation\_matrix = titanic\_df.corr()

print(correlation\_matrix)

1. Visualisasikan Korelasi

import seaborn as sns

import matplotlib.pyplot as plt

# Visualisasi matriks korelasi

sns.heatmap(correlation\_matrix, annot=True)

plt.title('Korelasi Antar Variabel')

plt.show()

1. Pair Plot

# Pair plot

sns.pairplot(correlation\_matrix)

plt.show()

1. Analisis

if correlation\_matrix ['Age'] ['Fare'] > 1.0:

print("\n Terdapat Korelasi positif")

else:

print("Tidak ada korelasi kuat")