UJIAN TENGAH SEMESTER (UTS)
MATA KULIAH DATA SCIENCE

Program Studi : Teknik Informatika
Semester/SKS : 7 / 3 SKS
Hari/Tanggal : - / -
Waktu : 2 Pekan
Sifat Ujian : Project Based
Dosen : Yulrio Brianorman

Capaian Pembelajaran yang Diukur:

CPMK	Indikator
CPMK-1	Mahasiswa memahami konsep dasar data science dan ruang lingkupnya
CPMK-2	Mahasiswa mampu melakukan pengumpulan, pembersihan, dan eksplorasi data
CPMK-3	Mahasiswa mampu menerapkan metode statistik dan visualisasi data

Petunjuk Pengerjaan:

1. Berdoalah sebelum mengerjakan project
2. Bentuk kelompok terdiri dari 3-4 mahasiswa
3. Kerjakan semua soal secara sistematis dan kolaboratif
4. Gunakan Python (Jupyter Notebook/Google Colab) untuk analisis data
5. Dataset: *Website Performance Dataset* dari Kaggle (`https://www.kaggle.com/datasets/maidasajid/website-performance-dataset`)
6. Buat video presentasi dengan durasi 12-15 menit (format MP4, min. 720p)
7. Deliverables yang harus dikumpulkan:
    - File Jupyter Notebook (.pdf)
    - Link Video presentasi kelompok
    - Slide presentasi (PowerPoint/PDF)
    - Laporan kontribusi individu (1-2 halaman per anggota)
8. Format penamaan file: `KelompokXX_NamaDataset_Item`
9. Deadline pengumpulan: 5 Desember 2025 pukul 23:59 WIB
10. Keterlambatan pengumpulan dikenakan penalty -5 poin per hari

---

BAGIAN A: Data Wrangling dan Exploratory Data Analysis (40 poin)

1. (10 points) **Data Wrangling**

   Lakukan pembersihan data pada *Website Performance Dataset*.

   (a) (4 points) Identifikasi dan tampilkan informasi tentang:

   - Jumlah baris dan kolom
   - Tipe data setiap kolom
   - Missing values pada setiap kolom
   - Duplicate data

   (b) (3 points) Tangani missing values dengan metode yang sesuai (mean imputation, median imputation, atau drop). Jelaskan alasan pemilihan metode untuk setiap variabel.

   (c) (3 points) Deteksi dan tangani outliers pada variabel `response_time` menggunakan metode IQR atau Z-score. Visualisasikan sebelum dan sesudah penanganan outliers menggunakan boxplot.

2. (15 points) **Analisis Statistik Deskriptif**

   Lakukan analisis statistik deskriptif pada dataset yang telah dibersihkan.

   (a) (5 points) Hitung dan tampilkan statistik deskriptif untuk variabel numerik berikut:

   - `response_time`: mean, median, std, min, max, Q1, Q3, P95
   - `throughput`: mean, median, std, min, max
   - `page_load_time`: mean, median, std, min, max

   (b) (5 points) Buat visualisasi distribusi untuk ketiga variabel di atas menggunakan:

   - Histogram dengan KDE (Kernel Density Estimation)
   - Boxplot

   (c) (5 points) Interpretasikan hasil statistik deskriptif dan visualisasi:

   - Apakah distribusi data normal atau skewed?
   - Identifikasi apakah terdapat long tail pada response time
   - Apa implikasi dari distribusi tersebut terhadap user experience?

3. (15 points) **Analisis Bivariate dan Multivariate**

   (a) (5 points) Buat analisis hubungan antara `response_time` dan `channel` (misalnya: Social Media, Email, Direct):

   - Hitung statistik deskriptif response time untuk setiap channel
   - Visualisasikan dengan boxplot atau violin plot
   - Identifikasi channel mana yang memiliki performa terbaik dan terburuk

   (b) (5 points) Analisis korelasi antar variabel numerik:

   - Hitung correlation matrix
   - Visualisasikan dengan heatmap
   - Identifikasi 3 pasang variabel dengan korelasi tertinggi

(c) (5 points) Buat scatter plot untuk mengeksplorasi hubungan antara:

- `page_load_time` vs `bounce_rate`
- `response_time` vs `conversion_rate`

Interpretasikan pola yang terlihat pada scatter plot.

BAGIAN B: Uji Statistik Hipotesis (40 poin)

4. (20 points) **Perbandingan Rata-Rata (T-Test Sederhana)**

Pilih 2 grup dari dataset untuk dibandingkan. Contoh opsi:

- Bandingkan response time untuk **2 jenis browser** yang berbeda (jika ada kolom browser)
- Bandingkan response time untuk **2 lokasi/region** yang berbeda (jika ada kolom location)
- Bandingkan response time untuk **2 kategori waktu** (peak hours vs non-peak hours)
- Atau buat kategori sendiri: response time rendah (`< median`) vs tinggi (`>= median`)

Catatan: Sesuaikan dengan kolom yang tersedia di dataset Anda

(a) (5 points) Siapkan data:

- Identifikasi kolom kategorikal yang akan digunakan untuk grouping
- Filter data untuk grup pertama dan grup kedua
- Hitung mean, median, dan standar deviasi response time masing-masing grup
- Tampilkan jumlah sampel di setiap grup

(b) (8 points) Lakukan T-Test:

- H_0: Rata-rata response time kedua grup sama
- H_1: Rata-rata response time kedua grup berbeda
- Gunakan fungsi: `from scipy.stats import ttest_ind`
- Code: `ttest_ind(group1_data, group2_data)`
- Catat hasil: t-statistic dan p-value

(c) (4 points) Visualisasi dengan boxplot atau violin plot untuk kedua grup

(d) (3 points) Kesimpulan:

- Jika p-value < 0.05: Ada perbedaan signifikan
- Jika p-value ≥ 0.05: Tidak ada perbedaan signifikan
- Grup mana yang memiliki performa lebih baik? Jelaskan!

5. (20 points) **Hubungan Antar Variabel (Correlation)**

Analisis hubungan antara dua variabel numerik di dataset. Contoh pasangan variabel:

- `response_time` vs `throughput`
- `page_size` vs `load_time`
- `number_of_requests` vs `response_time`

Pilih pasangan variabel yang ada di dataset Anda

(a) (5 points) Siapkan data:

- Pilih 2 variabel numerik yang ingin dianalisis hubungannya
- Pastikan tidak ada missing values
- Tampilkan statistik deskriptif keduanya (mean, median, std)

(b) (8 points) Hitung Korelasi:

- H_0: Tidak ada korelasi antara kedua variabel
- H_1: Ada korelasi antara kedua variabel
- Gunakan: `df[['var1', 'var2']].corr()`
- Atau: `from scipy.stats import pearsonr`
- Catat nilai correlation coefficient (r) dan p-value

(c) (4 points) Buat scatter plot dengan regression line menggunakan `seaborn.regplot()`

(d) (3 points) Kesimpulan:

- Apakah ada hubungan signifikan? (p-value < 0.05)
- Hubungan positif atau negatif?
- Kuat atau lemah? (lihat panduan di bawah)
- Apa implikasi dari temuan ini untuk optimasi website?

Panduan Mudah Interpretasi:

- P-value < 0.05 = Ada hubungan/perbedaan yang signifikan ✓
- P-value ≥ 0.05 = Tidak ada hubungan/perbedaan yang signifikan ×
- Correlation 0–0.3 = Lemah, 0.3–0.7 = Sedang, 0.7–1.0 = Kuat
- Correlation positif (+) = naik bersama, negatif (–) = berlawanan

BAGIAN C: Analisis Long Tail dan Business Insight (20 poin)

6. (10 points) **Analisis Long Tail Performance**

   (a) (4 points) Buat Cumulative Distribution Function (CDF) untuk variabel `response_time`.

   (b) (3 points) Identifikasi:

   - Persentase pengguna yang mengalami response time di atas 5 detik
   - Persentase pengguna yang mengalami response time di atas P95
   - Nilai response time pada P50, P90, P95, dan P99

   (c) (3 points) Jelaskan dampak long tail response time terhadap user experience dan potensi dampaknya terhadap business metrics.

7. (10 points) **Business Insight dan Rekomendasi**

   Berdasarkan hasil EDA dan uji statistik yang telah dilakukan, buat laporan insight dan rekomendasi.

(a) (6 points) Tuliskan minimal 5 insight penting yang ditemukan dari analisis data, mencakup:

- Insight tentang distribusi response time dan pola performa website
- Insight tentang perbedaan performa antar grup/kategori yang dianalisis
- Insight tentang hubungan antar variabel numerik (dari analisis korelasi)
- Insight tentang outliers dan extreme values dalam data
- Insight tentang faktor-faktor yang mempengaruhi performa website

(b) (4 points) Berikan minimal 3 rekomendasi strategis untuk meningkatkan website performance, disertai dengan justifikasi berdasarkan hasil analisis statistik dan visualisasi yang telah dibuat.

Catatan Penilaian:

- Kode Python yang rapi dan terstruktur: 10% dari total nilai
- Visualisasi yang informatif dan mudah dibaca: 15% dari total nilai
- Interpretasi dan analisis yang mendalam: 25% dari total nilai
- Ketepatan metodologi dan hasil uji statistik: 50% dari total nilai

— SELAMAT MENGERJAKAN —