1. Jika algoritma K-Means menghasilkan nilai silhouette score rendah (0.3) meskipun elbow method menunjukkan K=5 sebagai optimal pada dataset ini, faktor apa yang menyebabkan inkonsistensi ini? Bagaimana strategi validasi alternatif (misal: analisis gap statistic atau validasi stabilitas cluster via bootstrapping) dapat mengatasi masalah ini, dan mengapa distribusi data non-spherical menjadi akar masalahnya?

   Jika data berbentuk tidak beraturan, maka centroid-based clustering gagal mengelompokkannya dengan baik, meski inertia pada elbow method menurun signifikan. Overlap antar cluster juga menurunkan silhouette score. Centroid based model bermasalah karena ia berdasarkan jarak Euclidean atau jarak lain, dan akan struggle untuk clustering yang Panjang atau melengkung.
   Alternatif berupa gap statistic yaitu membandingkan total intertia dari data asli dengan intertia dari data acak, jika data asli tidak jauh lebih baik dari adata acak, maka cluster tidak signifikan. Alternatif lain berupa bootstrapping, melatih ulang KMeans dan melihat konsistensi cluster.

2. Dalam dataset dengan campuran fitur numerik (Quantity, UnitPrice) dan kategorikal high-cardinality (Description), metode preprocessing apa yang efektif untuk menyelaraskan skala dan merepresentasikan fitur teks sebelum clustering? Jelaskan risiko menggunakan One-Hot Encoding untuk Description, dan mengapa teknik seperti TF-IDF atau embedding berdimensi rendah (UMAP) lebih robust untuk mempertahankan struktur cluster!

   Untuk fitur numerik bisa langsung discaling, di notebook saya namun membuat fitur baru terlebih dahulu untuk mengelompokkan pengeluaran berdasarkan customer. Untuk description, seharusnya fiturnya tidak terpakai karena sudah ada fitur StockCode. Di notebook saya StockCode juga tidak terpakai karena saya akan clustering berdasarkan customer.
   Namun untuk menjawab pertanyaan, One-Hot encoding akan membuat fitur banyak sesusai jumlah kategori uniknya, maka akan banyak noise. TF-IDF lebih baik karena berdasarkan kata-kata yang muncul, atau Word2Vec juga dapat digunakan untuk mengubah deskripsi menjadi numerik.

3. Hasil clustering dengan DBSCAN sangat sensitif terhadap parameter epsilon—bagaimana menentukan nilai optimal epsilon secara adaptif untuk memisahkan cluster padat dari noise pada data transaksi yang tidak seimbang (misal: 90% pelanggan dari UK)? Jelaskan peran k-distance graph dan kuartil ke-3 dalam automasi parameter, serta mengapa MinPts harus disesuaikan berdasarkan kerapatan regional!

   Epsilon menentukan jarak maksimum Dimana sebuah titik masih dianggap tentangga dari titiknya. Epsilon terlalu kecil, maka Sebagian besar titik akan dianggap noise, dan epsilon terlalu besar maka cluster akan digabungkan menjadi 1.

K-distance graph dapat menentukan Epsilon optimal dengan mencari elbow di plot k-distance.
Q3 memberikan nilai epsilon yang adaptif yang cukup besar untuk mengcover cluster padat dan cukup kecil untuk tidak menyerap noise.
MinPts disesuaikan berdasarkan kerapatan regional misalnya pada UK yang transaksinya lebih padat, bisa gunakan MinPts lebih tinggi, dan untuk negara lain harus disesuaikan menggunakan MinPts lebih rendah.

4. Jika analisis post-clustering mengungkapkan overlap signifikan antara cluster "high-value customers" dan "bulk buyers" berdasarkan total pengeluaran, bagaimana teknik semi-supervised (contoh: constrained clustering) atau integrasi metric learning (Mahalanobis distance) dapat memperbaiki pemisahan cluster? Jelaskan tantangan dalam mempertahankan interpretabilitas bisnis saat menggunakan pendekatan non-Euclidean!

   Contained Clustering menerapkan dua aturan yaitu must-link dan cannot-link, contohnya customer A dan B diketahui sebagai bulk buyer jadi must link, namun customer C high-value customer jadi cannot-link dengan A atau B. Dengan ini kitab isa mengarahkan model agar mengerti ke perbedaan yang tidak terlihat hanya dari angka total pembelian.
   Metrik learning untuk mempelajari jarak yang relevan secara bisnis, bukan sekadar jarak spasial/Euclidian. Mahalanobis Distance mengukur jarak dengan mempertimbangkan korelasi antar fitur dan skala fitur.
   Tantangannya berupa dimensi tidak langsung terlihat karena tidak mengikuti sumbu asli, serta lebih kompleks.

5. Bagaimana merancang temporal features dari InvoiceDate (misal: hari dalam seminggu, jam pembelian) untuk mengidentifikasi pola pembelian periodik (seperti transaksi pagi vs. malam)? Jelaskan risiko data leakage jika menggunakan agregasi temporal (misal: rata-rata pembelian bulanan) tanpa time-based cross-validation, dan mengapa lag features (pembelian 7 hari sebelumnya) dapat memperkenalkan noise pada cluster!

   Untuk mendapatkan fitur temporal dapat dipisahkan InvoiceDate menjadi fitur-fitur sendiri seperti tahun, bulan, minggu (7 hari), hari, jam, dll. Resiko data leakage seperti ketika menghitung rata-rata pembelian bulanan tidak sengaja akan menggunakan data maret untuk prediksi data februari. Clustering juga berpengaruh karena hasil cluter tidak merfefleksikan perilaku yang benar ada. Time based cross validation membagi data berdasarkan waktu dengan urut.
   Lag feature dapat bermasalah karena clustering tidak tau arah waktu,