Persiapan Data

Dataset yang digunakan adalah data penjualan produk (data_penjualan.csv). Dataset ini berisi data transaksi yang mencakup informasi kuantitatif seperti Jumlah Order, Harga, dan Total. Dataset ini cocok untuk analisis karena memiliki data numerik, kategorikal, dan bisa juga teks (misalnya nama produk, kategori, atau wilayah penjualan). Dataset diambil dari kaggle dan dapat diakses pada link berikut: https://www.kaggle.com/datasets/jabirmuktabir/data-penjualan-produk-cetakan

1. Langkah 1: Analisis, Telaah, dan Validasi Data

   Jumlah baris dan kolom diperiksa, serta tipe data diidentifikasi dengan df.info(). Berikut fitur-fitur yang ditemukan pada dataset:

   - Tanggal → datetime
   - Jenis Produk → kategorik
   - Jumlah Order → numerik
   - Harga → numerik
   - Total → numerik

   Statistik deskriptif (mean, min, max, dll.) ditampilkan untuk kolom numerik.

   Tidak ditemukan nilai kosong atau missing values.

   Outlier terdeteksi di:

   - Jumlah Order: 95 entri
   - Harga: 33 entri
   - Total: 69 entri

   Outlier yang terdeteksi pada kolom Jumlah Order (95 entri), Harga (33 entri), dan Total (69 entri) tidak dihapus karena berpotensi merepresentasikan kasus nyata, seperti pembelian dalam jumlah besar, harga promo khusus, atau diskon ekstrem. Menghapus outlier justru berisiko menghilangkan informasi penting yang relevan dengan analisis bisnis. Oleh karena itu, outlier tetap dipertahankan dan akan diperhatikan secara khusus dalam tahap analisis selanjutnya.

2. Langkah 2: Strategi Pembersihan Data

   Berdasarkan hasil telaah, strategi pembersihan data yang ditetapkan adalah tidak melakukan penghapusan data karena tidak ditemukan missing value. Outlier dipertahankan karena memiliki kemungkinan relevan secara bisnis. Fokus utama pembersihan diarahkan pada konsistensi nilai numerik dan standarisasi nama kolom agar data lebih mudah dianalisis.

3. Langkah 3: Koreksi Data Kotor

Tahap berikutnya adalah mengoreksi data yang kotor. Langkah pertama dilakukan dengan membersihkan nama kolom dari spasi yang tidak diperlukan. Selanjutnya, kolom Jumlah Order, Harga, dan Total dikonversi menjadi tipe data numerik untuk memastikan konsistensi. Selain itu, dilakukan perhitungan ulang pada kolom Total berdasarkan perkalian antara Jumlah Order dan Harga, yang kemudian disimpan dalam kolom baru bernama Total_Recalc. Hasil verifikasi menunjukkan tidak ada data yang memiliki nilai total tidak sesuai dengan hasil perhitungan ulang, yang menandakan tidak adanya kesalahan entri data.

4. Langkah 4: Melakukan Transformasi Data

   Untuk mempersiapkan data agar lebih siap digunakan dalam analisis, dilakukan beberapa tahap transformasi. Pertama, kolom Tanggal dikonversi menjadi format datetime agar dapat dimanfaatkan dalam analisis berbasis waktu. Selanjutnya, kolom numerik seperti Jumlah Order, Harga, dan Total dikonversi secara eksplisit ke tipe data numerik untuk memastikan konsistensi format. Setelah itu, dilakukan rekayasa fitur dari kolom tanggal dengan mengekstraksi informasi Tahun, Bulan, Hari, serta nama hari (HariMinggu). Selain itu, ditambahkan pula kolom Periode dalam format YYYY-MM yang siap digunakan untuk analisis tren bulanan. Transformasi ini membuat dataset lebih kaya dengan informasi temporal, terstruktur dengan baik, dan memudahkan analisis pola penjualan berdasarkan waktu, misalnya tren musiman, pola mingguan, atau perbandingan antar periode.

5. Langkah 5: Membuat Dokumentasi Konstruksi Data

   Proses konstruksi data didokumentasikan secara sistematis mulai dari tahap cleaning, koreksi, hingga transformasi. Teknik yang digunakan meliputi pembersihan nama kolom, konversi tipe data numerik, validasi nilai melalui perhitungan ulang, serta penerapan normalisasi dan encoding. Dampak dari proses ini adalah meningkatnya kualitas data yang lebih bersih, konsisten, dan reliabel, sehingga dapat meningkatkan keakuratan hasil analisis serta memudahkan proses pemodelan pada tahap selanjutnya.

6. Langkah 6:Melakukan Pelabelan Data

   Pelabelan data dilakukan dengan membagi nilai pada kolom Total ke dalam tiga kategori berdasarkan kuartil distribusinya. Teknik yang digunakan adalah pd.qcut, yaitu metode yang secara otomatis membagi data menjadi sejumlah interval dengan jumlah anggota yang seimbang. Pada kasus ini, Total dibagi ke dalam tiga kelas, yaitu Rendah, Sedang, dan Tinggi. Dengan pendekatan ini, data numerik yang bersifat kontinyu dapat diubah menjadi data kategorikal yang lebih mudah dianalisis, misalnya untuk tujuan segmentasi pelanggan atau analisis perbandingan antar kelompok. Hasil pelabelan menghasilkan

distribusi yang seimbang di tiap kategori karena berbasis pada kuartil, sehingga mengurangi bias akibat ketidakseimbangan jumlah data antar kelas.

7. Langkah 7: Membuat Laporan Hasil Pelabelan Data

   Distribusi Label

   Dataset penjualan telah diberi label kategori Total menggunakan dua pendekatan:

   - Menggunakan qcut (berdasarkan kuantil):

     Data terbagi menjadi tiga kategori yang relatif seimbang, masing-masing ±33% dari total data.

     - Rendah: ±33%
     - Sedang: ±33%
     - Tinggi: ±33%

     Keseimbangan ini terjadi karena qcut membagi data berdasarkan posisi relatif (persentil), bukan nilai mutlak. Dengan demikian, jumlah entri di setiap kategori hampir sama, meskipun nilai penjualan antar kategori bisa sangat berbeda.
   - Menggunakan cut (berdasarkan nilai asli dengan batas maksimal 10 juta):
     - Rendah: (0 – 3,3 juta) → proporsi terbesar, karena mayoritas transaksi bernilai rendah.
     - Sedang: (3,3 – 6,6 juta) → proporsi sedang, jumlah transaksi cukup signifikan.
     - Tinggi: (> 6,6 juta) → proporsi terkecil, hanya sebagian kecil transaksi yang bernilai tinggi.

     Distribusi tidak seimbang karena pembagian kategori mengikuti nilai absolut dari data, sehingga lebih mencerminkan kondisi riil bisnis.

   Evaluasi Proses Pelabelan

   Kelebihan:

   - Versi qcut:
     - Memberikan distribusi kelas yang merata, sehingga cocok untuk model machine learning supervised yang sensitif terhadap ketidakseimbangan kelas.
     - Meminimalkan risiko bias terhadap kelas mayoritas.

   - Versi cut:
     - Lebih representatif terhadap realitas bisnis, karena kategori dibentuk dari nilai transaksi aktual.

- Mudah dipahami oleh stakeholder non-teknis (misalnya: transaksi > 6,6 juta langsung dianggap “Tinggi”).

Tantangan:

Pada qcut, batas antar kategori ditentukan statistik, bukan logika bisnis. Misalnya, transaksi Rp 3,5 juta bisa dianggap “Tinggi” kalau posisinya masuk persentil atas, padahal dalam praktik bisnis mungkin masih dianggap rendah. Pada cut, distribusi kelas bisa timpang (misalnya 70% “Rendah”), sehingga berpotensi menyebabkan masalah pada model klasifikasi (class imbalance). Nilai outlier tetap berpengaruh pada pembagian kategori, khususnya di kelas “Tinggi”.

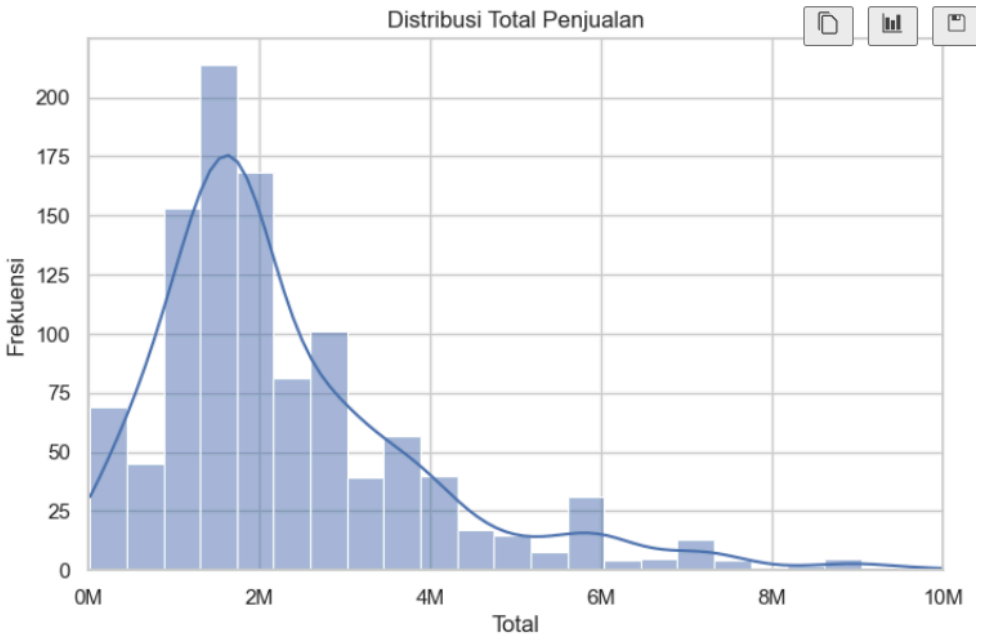
Rekomendasi Perbaikan

- Validasi dengan domain knowledge:
  Mendiskusikan dengan pihak bisnis untuk menentukan ambang batas yang lebih realistis, misalnya:
  - Rendah < 2 juta
  - Sedang 2–5 juta
  - Tinggi > 5 juta
- Kombinasi metode:
  Gunakan qcut untuk keperluan analisis eksploratif dan machine learning, tapi gunakan cut (nilai absolut) untuk laporan bisnis.
- Penanganan outlier:
  Terapkan metode winsorization atau trimming agar distribusi kategori tidak terlalu dipengaruhi transaksi ekstrem.
- Feature engineering tambahan:
  Buat label gabungan yang memperhitungkan variabel lain (misalnya Total + Jumlah Order) untuk menghasilkan kategori yang lebih informatif.

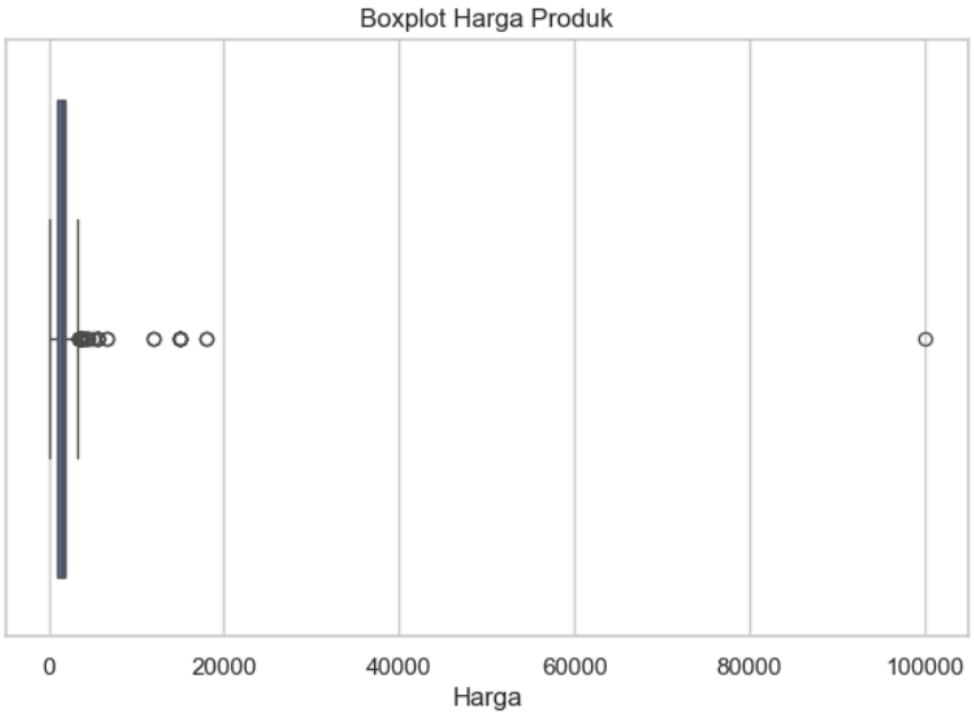
8. <u>Langkah 8: Visualisasi Data</u>

   Setelah melalui tahap pembersihan, transformasi, dan rekayasa fitur, data divisualisasikan untuk memahami distribusi serta pola yang terkandung di dalamnya. Beberapa bentuk visualisasi yang digunakan adalah:

   1. Histogram digunakan untuk melihat distribusi numerik seperti kolom Total, Jumlah Order, dan Harga. Histogram ini menunjukkan bahwa mayoritas transaksi memiliki nilai total relatif kecil, sedangkan transaksi dengan nilai tinggi jumlahnya sedikit, menandakan adanya kemiringan (skewness) ke arah kanan.

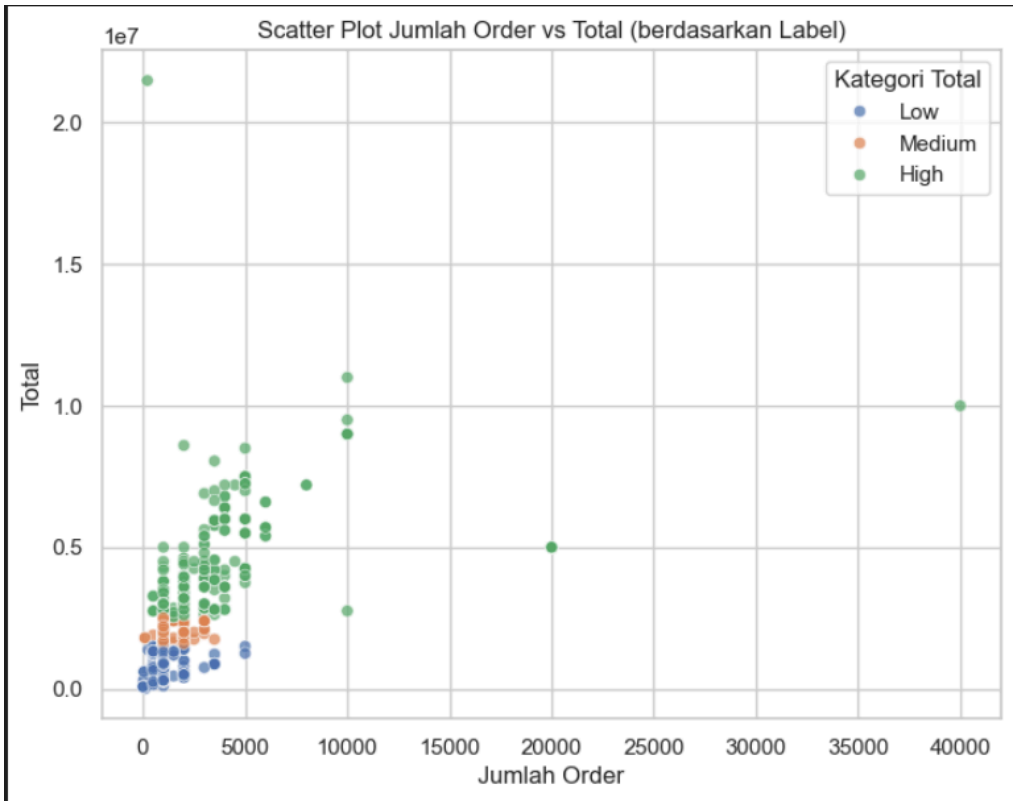


2. Box Plot digunakan untuk mendeteksi keberadaan outlier pada variabel numerik, khususnya Total. Dari box plot terlihat bahwa terdapat sejumlah nilai ekstrem (outlier) yang jauh lebih tinggi dibandingkan mayoritas data. Outlier ini bisa memberikan insight tentang adanya transaksi besar, tetapi juga berpotensi memengaruhi analisis statistik.

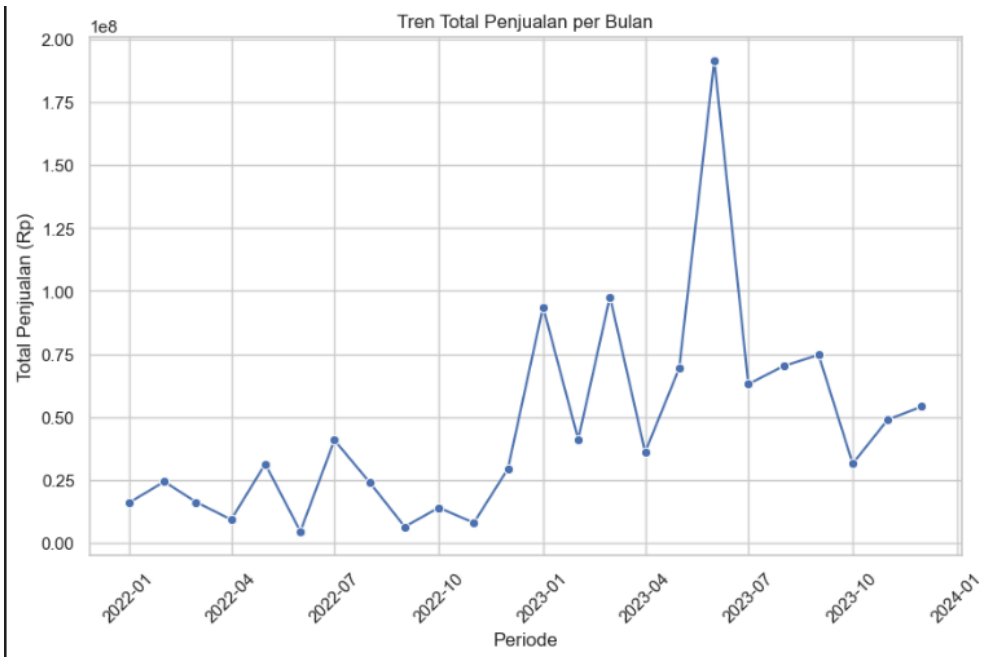


3. Scatter Plot → memvisualisasikan hubungan antara Jumlah Order dengan Total. Terlihat korelasi positif: semakin banyak jumlah order, semakin tinggi nilai total transaksi. Namun, ada beberapa titik menyebar jauh dari pola utama,

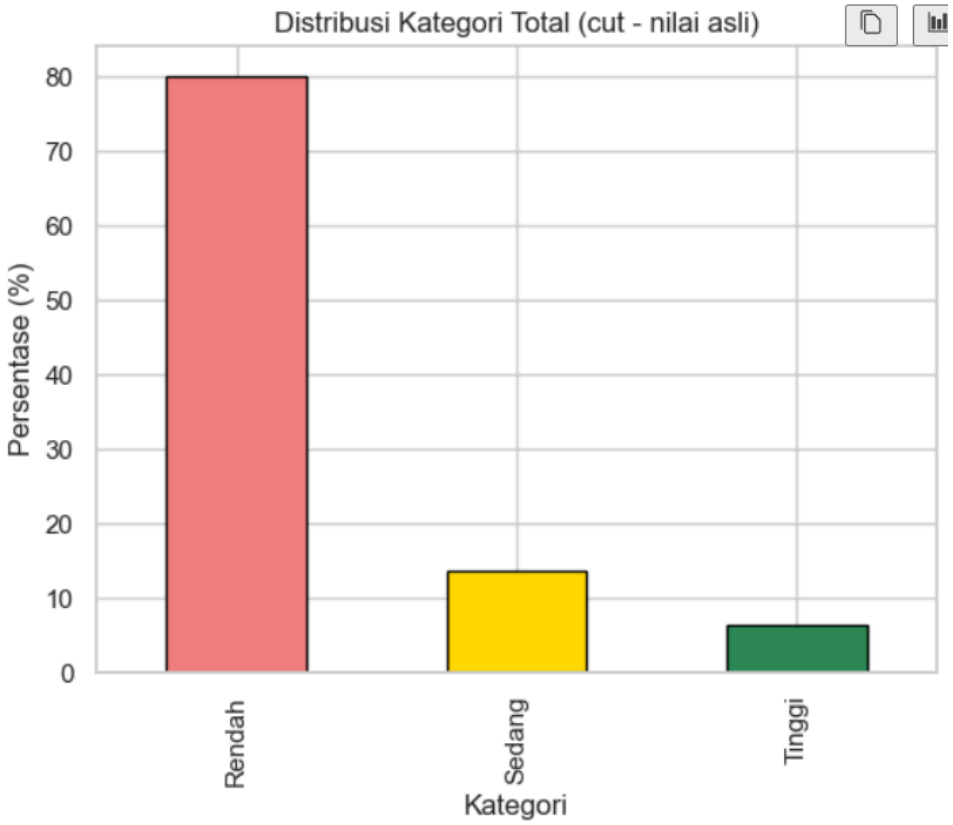
mengindikasikan transaksi tidak biasa (misalnya order kecil tapi total tinggi karena harga per unit mahal).



4. Grafik tren penjualan per bulan yang dihasilkan dari line chart tersebut berfungsi untuk melihat dinamika total penjualan sepanjang waktu. Data awal terlebih dahulu direkayasa dengan membuat kolom Periode (gabungan Tahun dan Bulan), kemudian dilakukan agregasi sehingga diperoleh jumlah penjualan per bulan. Hasil agregasi inilah yang divisualisasikan menggunakan line chart dengan tambahan marker di setiap titik bulan agar perubahan lebih jelas terlihat. Dari grafik, dapat diamati pola naik-turun omzet yang mencerminkan fluktuasi transaksi pada periode tertentu.



5. Bar Chart (Distribusi Label) → memvisualisasikan hasil pelabelan kategori Rendah, Sedang, dan Tinggi. Grafik ini membantu memahami proporsi kelas yang terbentuk, baik dari metode qcut (seimbang) maupun cut (mengikuti nilai asli). Visualisasi ini memperjelas struktur data, distribusinya, serta potensi masalah seperti ketidakseimbangan kelas dan outlier.



9. Langkah 9: Evaluasi dan Dokumentasi

   Secara keseluruhan, proses konstruksi data telah dilakukan mulai dari pemahaman data mentah, pembersihan nilai hilang dan salah format, transformasi numerik maupun kategorikal, rekayasa fitur dari tanggal, pelabelan berbasis kuantil maupun nilai asli, hingga tahap visualisasi.

   Hasil yang dicapai:

   - Dataset menjadi lebih bersih dan terstruktur. Kolom yang semula memiliki format tidak konsisten (seperti tanggal dan angka) berhasil dikonversi.
   - Fitur tambahan seperti Tahun, Bulan, Hari, dan HariMinggu meningkatkan granularitas analisis.
   - Pelabelan data memberikan cara baru untuk membedakan transaksi berdasarkan skala nilai.
   - Visualisasi menunjukkan distribusi data yang tidak merata serta keberadaan outlier.

   Rekomendasi tindak lanjut:

   - Validasi domain knowledge: Lakukan diskusi dengan stakeholder untuk menentukan batas kategorisasi yang lebih relevan dengan konteks bisnis.
   - Tangani outlier: Pertimbangkan metode winsorization atau log transformation agar data lebih stabil digunakan pada model prediktif.
   - Analisis lanjutan: Gunakan dataset yang sudah bersih untuk membangun model prediksi (misalnya klasifikasi atau regresi).
   - Automasi pipeline: Dokumentasikan dan simpan semua langkah pembersihan serta transformasi agar dapat diulang (reproducible) untuk dataset baru.