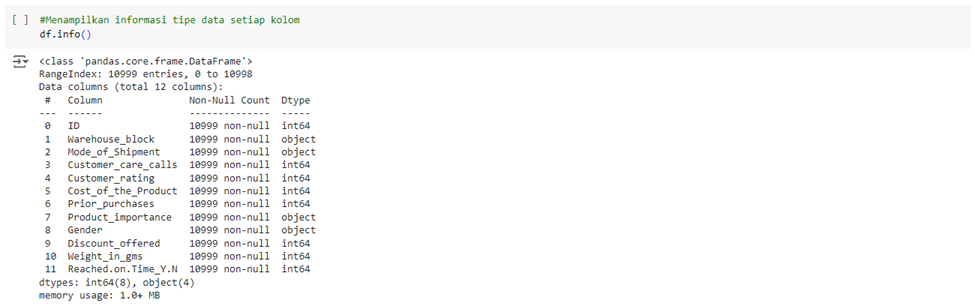
**Homework - EDA**

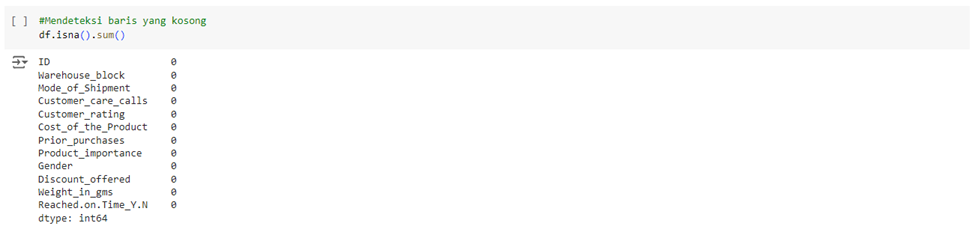
**(Kelompok 1 - SA7)**

|  |
| --- |
| **Database**: <https://www.kaggle.com/datasets/prachi13/customer-analytics/data> |

1. **Descriptive Statistics**
2. Kolom dan Tipe Data

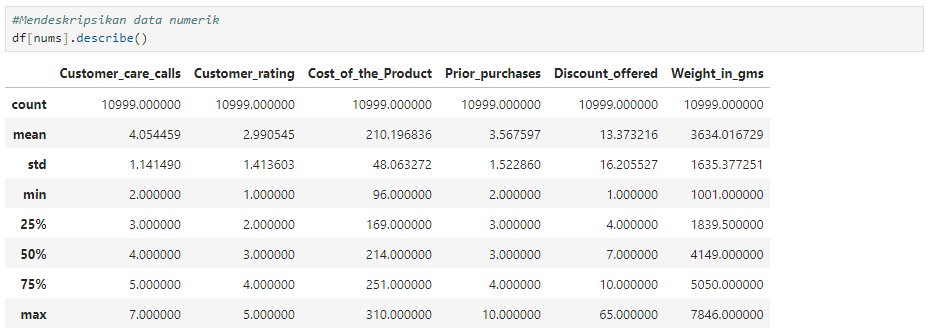
Berdasarkan hasil pengecekan, semua tipe data telah sesuai dengan yang diharapkan. Oleh karena itu, proses selanjutnya dapat dilanjutkan.

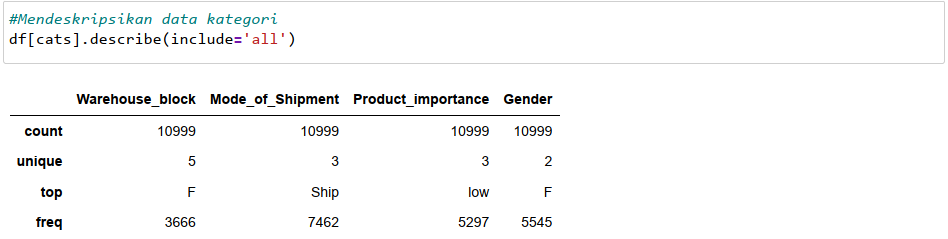
1. Handle Missing Value

Berdasarkan hasil pemeriksaan terhadap keberadaan nilai kosong, dapat disimpulkan bahwa database yang digunakan tidak mengandung missing values pada seluruh kolomnya. Dengan demikian, proses penanganan nilai kosong dapat diabaikan.

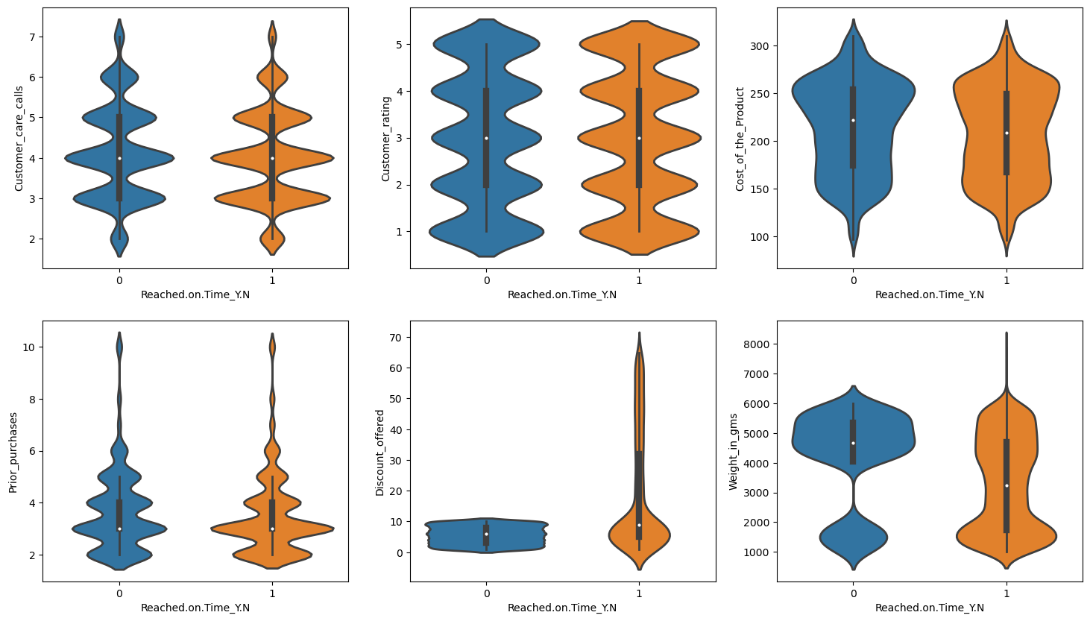
1. Nilai Summary

Tidak ditemukan adanya nilai data yang menyimpang. Semua nilai data telah sesuai dengan standar yang ditetapkan.



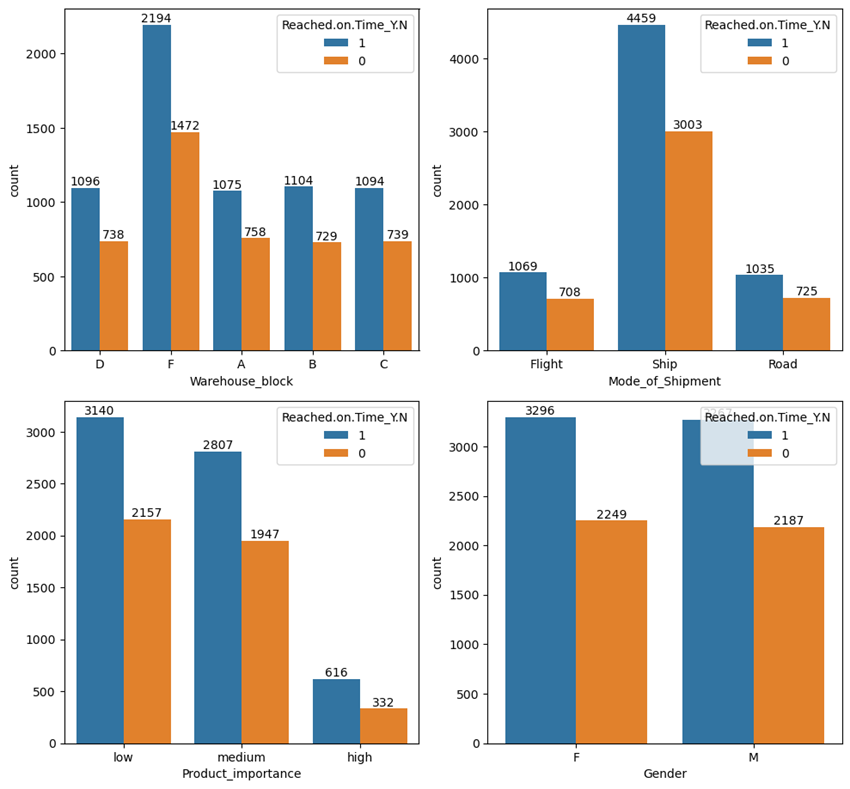


1. **Univariate Analysis**
2. Violin Plot



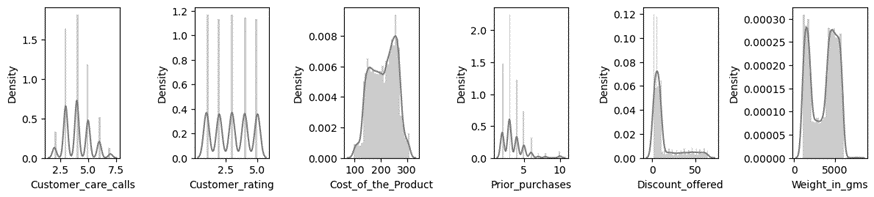
Analisis distribusi menggunakan violin plot menunjukkan bahwa:

1. Customer Care Calls: Distribusi panggilan layanan pelanggan serupa untuk kedua kategori, menunjukkan bahwa status pengiriman tepat waktu tidak secara signifikan memengaruhi jumlah panggilan layanan pelanggan.
2. Customer Rating: Peringkat pelanggan tampaknya sedikit lebih tinggi untuk kategori “Reached on Time\_YN” yaitu 0 (pengiriman tepat waktu) dibandingkan dengan kategori 1 (pengiriman terlambat). Namun, tumpang tindih antara distribusi tersebut menunjukkan bahwa masih ada sebagian besar pelanggan yang menilai pengiriman sangat tinggi meskipun terlambat.
3. Cost of the Product: Biaya produk tampaknya sedikit lebih tinggi untuk kategori “Reached on Time\_YN” 1 (pengiriman terlambat). Hal ini mungkin menunjukkan bahwa produk yang lebih mahal lebih mungkin untuk dikirim terlambat.
4. Prior Purchases: Distribusi pembelian sebelumnya serupa untuk kedua kategori, menunjukkan bahwa jumlah pembelian sebelumnya tidak secara signifikan mempengaruhi pengiriman tepat waktu.
5. Prior Purchases: Distribusi diskon yang ditawarkan sedikit lebih luas untuk kategori “Reached on Time\_YN” sebesar 0 (pengiriman tepat waktu).
6. Weight in Grams: Distribusi berat dalam gram sedikit condong ke arah nilai yang lebih tinggi untuk kategori “Reached on Time\_YN” 1 (pengiriman terlambat). Hal ini mungkin menunjukkan bahwa barang yang lebih berat lebih mungkin untuk dikirim terlambat.
7. Bar Plot

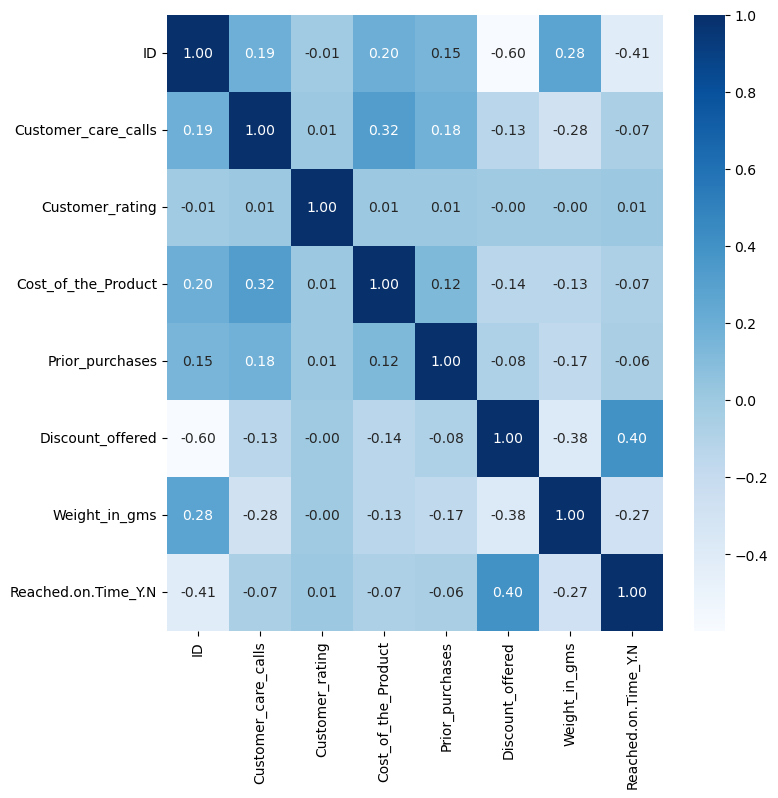


Analisis bar plot menunjukkan bahwa:

1. Warehouse Block: Distribusi blok gudang relatif seimbang di kedua kategori, menunjukkan bahwa blok gudang tidak secara signifikan memengaruhi ketepatan waktu pengiriman.
2. Mode of Shipment: Moda pengiriman penerbangan dan kapal lebih umum untuk pengiriman tepat waktu, sedangkan jalan darat lebih umum untuk pengiriman yang terlambat.
3. Product Importance: Produk yang diklasifikasikan sebagai produk dengan tingkat kepentingan “medium” atau “high” lebih mungkin dikirim tepat waktu, sedangkan produk yang diklasifikasikan sebagai produk dengan tingkat kepentingan “low” lebih mungkin dikirim terlambat.
4. Gender: Tidak ada perbedaan yang signifikan dalam distribusi jenis kelamin antara pengiriman yang tepat waktu dan terlambat, yang menunjukkan bahwa jenis kelamin bukanlah faktor yang mempengaruhi pengiriman tepat waktu.
5. Density Plot

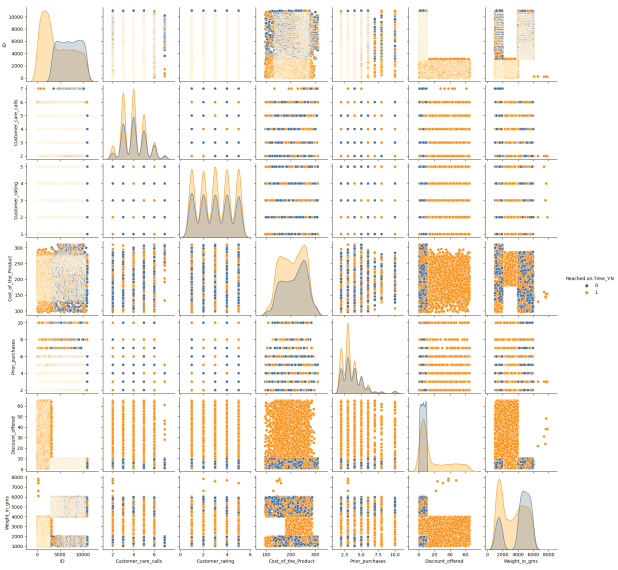
Data e-commerce ini menunjukkan distribusi yang tidak normal, terutama pada variabel 'prior\_purchases', 'discount\_offered', dan 'weight\_in\_gms' yang cenderung miring ke kanan (positively skewed). Kondisi ini dapat mengakibatkan uji statistik yang tidak valid, model Machine Learning yang overfitting pada data pelatihan, serta kesimpulan yang bias jika hanya bergantung pada rata-rata sebagai ukuran pusat data.

1. **Multivariate Analysis**
2. Heatmap



Tidak Ada Korelasi Sempurna: Fakta bahwa tidak ada korelasi yang nilainya persis 0 menunjukkan bahwa semua variabel memiliki hubungan satu sama lain, meskipun mungkin sangat lemah.

1. Pairplot



Data e-commerce ini memiliki distribusi tidak normal khususnya prior\_purchase,discount\_offered dan weight\_in\_gm condong skewed ke kanan. Hal ini akan berdampak pada uji statistik yang tidak valid, algoritma Machine Learning overfit dengan data latih, menyebabkan kesimpulan yang salah jika hanya bergantung pada rata-rata, menyebabkan pemisahan yang tidak proporsional di berbagai kelas histogram, dan membuat analisis visual menjadi menyesatkan.

Data e-commerce ini masih terdapat data outlier yang dapat mengganggu statistika deskriptif, pemodelan statistik, pengaruh pada Algoritma Machine Learning seperti overfitting dan keakuratannya hingga penarikan kesimpulan yang kurang tepat.

1. **Business Insight**

|  |  |  |
| --- | --- | --- |
| **Visualisasi Business Insight** | **(1) Mode Pengiriman dan Ketepatan Waktu** | **(2) Ketepatan Waktu Pengiriman Berdasarkan Kategori Berat** |
| **(3) Rata-rata Diskon Berdasarkan Mode Pengiriman** | **(4) Peluang untuk Mengoptimalkan Pengiriman Berdasarkan Kategori Berat** | **(5) Distribusi Pelanggan Berdasarkan Mode Pengiriman** |

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Business Insight** | **Insight** | **Rekomendasi** |
| 1. | Mode Pengiriman dan Ketepatan Waktu | Mode pengiriman Flight dan Ship lebih andal dalam pengiriman tepat waktu dibandingkan mode Road. Mode pengiriman via darat cenderung memiliki lebih banyak keterlambatan. | Prioritaskan mode pengiriman yang lebih andal untuk barang-barang penting atau dengan margin tinggi. Untuk mode pengiriman yang sering mengalami keterlambatan, tinjau ulang mitra logistik atau optimalkan rute pengiriman. |
| 2. | Ketepatan Waktu Pengiriman Berdasarkan Kategori Berat | Produk dalam kategori berat rendah dan menengah lebih cenderung dikirimkan tepat waktu, sementara produk dalam kategori berat tinggi sering kali terlambat. | Tingkatkan efisiensi logistik untuk produk dengan berat tinggi. Ini bisa dilakukan dengan mencari alternatif metode pengiriman yang lebih cepat atau bekerjasama dengan mitra logistik yang memiliki kapasitas untuk menangani barang berat secara lebih efisien. |
| 3. | Rata-rata Diskon Berdasarkan Mode Pengiriman | Meskipun diskon diberikan, ini tidak berdampak signifikan terhadap ketepatan waktu pengiriman. Hal ini menunjukkan bahwa diskon mungkin tidak secara langsung berpengaruh pada efisiensi pengiriman. | Fokuslah untuk meningkatkan pengalaman pengiriman yang konsisten, bukan hanya memberikan diskon. Memberikan insentif pada mitra pengiriman untuk memastikan ketepatan waktu bisa meningkatkan kepuasan pelanggan secara keseluruhan. |
| 4. | Peluang untuk Mengoptimalkan Pengiriman Berdasarkan Kategori Berat | Produk dengan berat tinggi menghadapi tantangan dari sisi biaya dan ketepatan waktu. Namun, jika pengiriman untuk kategori berat ini bisa dioptimalkan, ada potensi untuk meningkatkan margin penjualan dari produk yang lebih mahal dan lebih berat. | Berinvestasi dalam infrastruktur logistik untuk produk berat, seperti gudang yang lebih dekat dengan pelanggan atau memilih mitra pengiriman khusus untuk produk berat, bisa meningkatkan efisiensi operasional. |
| 5. | Distribusi Pelanggan Berdasarkan Mode Pengiriman | Setiap mode pengiriman mungkin memiliki segmen pelanggan yang berbeda. Misalnya, pelanggan yang memilih pengiriman via Flight mungkin lebih mengutamakan kecepatan daripada biaya, sedangkan pelanggan yang memilih pengiriman via Road mungkin lebih sensitif terhadap biaya. | Sesuaikan penawaran promosi dengan mode pengiriman yang dipilih pelanggan. Untuk pelanggan yang memilih pengiriman cepat, tawarkan layanan premium. Untuk pelanggan yang lebih peduli terhadap biaya, pertimbangkan menawarkan diskon atau paket hemat. |

1. **Git**

Link: