Homework EDA Stage 1

Kelompok 6 - dataset{'sat_set'}
Data: E-commerce Shipping Data
Ketua: Refanie FS
Anggota:

1. Nur Cahyanti
2. Handika
3. Indra Laksana
4. Fajar Nurdiono
5. Utlia Rahma

Outline

1. Descriptive Analytics
2. Univariate Analysis
3. Multivariate Analysis
4. Business Insights

[1] Descriptive Statistics

	ID	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
count	10999.00000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000
mean	5500.00000	4.054459	2.990545	210.196836	3.567597	13.373216	3634.016729	0.596691
std	3175.28214	1.141490	1.413603	48.063272	1.522860	16.205527	1635.377251	0.490584
min	1.00000	2.000000	1.000000	96.000000	2.000000	1.000000	1001.000000	0.000000
25%	2750.50000	3.000000	2.000000	169.000000	3.000000	4.000000	1839.500000	0.000000
50%	5500.00000	4.000000	3.000000	214.000000	3.000000	7.000000	4149.000000	1.000000
75%	8249.50000	5.000000	4.000000	251.000000	4.000000	10.000000	5050.000000	1.000000
max	10999.00000	7.000000	5.000000	310.000000	10.000000	65.000000	7846.000000	1.000000

	Warehouse_block	Mode_of_Shipment	Product_importance	Gender
count	10999	10999	10999	10999
unique	5	3	3	2
top	F	Ship	low	F
freq	3666	7462	5297	5545

```
Value count kolom Warehouse_block:
F    3666
D    1834
A    1833
B    1833
C    1833
Name: Warehouse_block, dtype: int64

Value count kolom Mode_of_Shipment:
Ship      7462
Flight    1777
Road      1760
Name: Mode_of_Shipment, dtype: int64
```

```
Value count kolom Product_importance:
low       5297
medium    4754
high       948
Name: Product_importance, dtype: int64

Value count kolom Gender:
F    5545
M    5454
Name: Gender, dtype: int64
```

[1] Descriptive Statistics

Apakah ada kolom dengan tipe data kurang sesuai, atau nama kolom dan isinya kurang sesuai?

Semua tipe data sudah sesuai. Nama kolom dan isinya sudah sesuai, kecuali kolom Warehouse_block memiliki data A, B, C, D, F dan bukan A, B, C, D, E sesuai keterangan dari Kaggle.

Apakah ada kolom yang memiliki nilai kosong? Jika ada, apa saja?

Tidak ada data yang memiliki nilai kosong.

Apakah ada kolom yang memiliki nilai summary agak aneh? (min/mean/median/max/unique/top/freq)

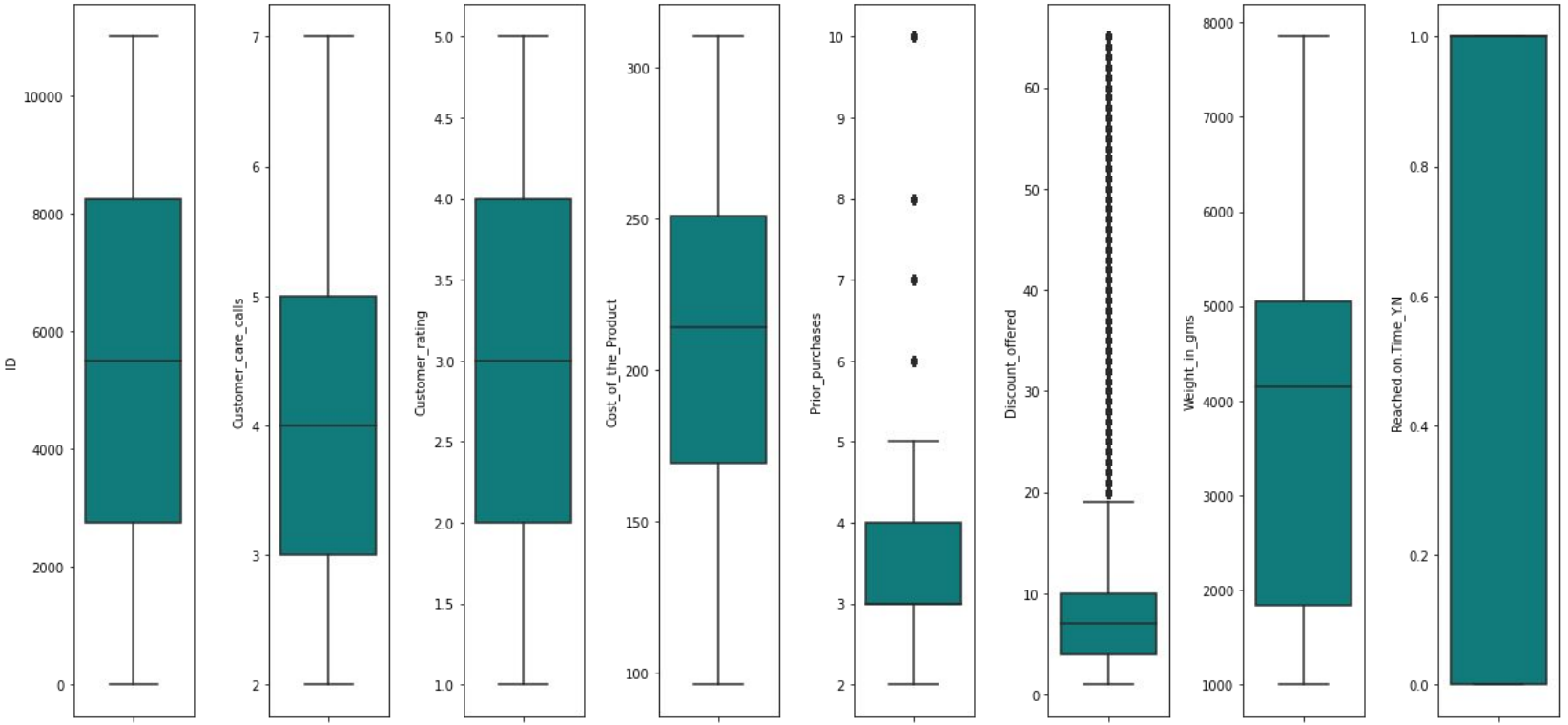
Pada data **Prior_purchases** terdapat perbedaan yang cukup besar pada mean dan mediannya. Nilai maksimumnya juga sangat besar, sehingga data ini pasti memiliki outliers.

Pada data **Discount_offered** juga terdapat perbedaan yang besar antara mean dan mediannya. Begitupun nilai maximum nya sangat besar, sehingga data ini juga sudah pasti memiliki outliers.

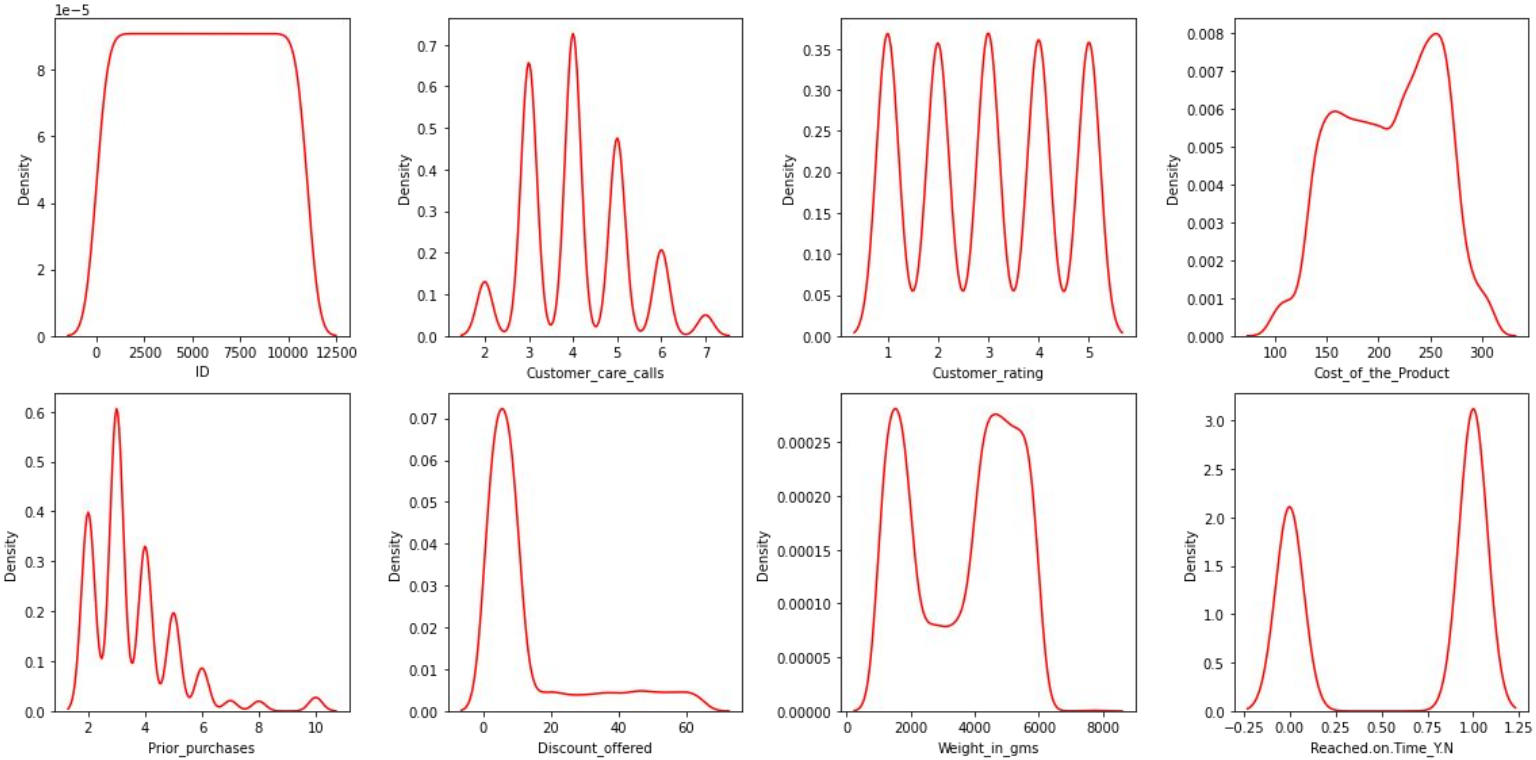
Data **Mode_of_Shipment** memiliki data 'Ship' yang sangat besar dibandingkan dengan data yang lain.

Sedangkan pada data **Product_importance** memiliki data 'high' yang jauh lebih kecil daripada yang lainnya.

[2] Univariate Analysis: Data Numerik



[2] Univariate Analysis: Data Numerik

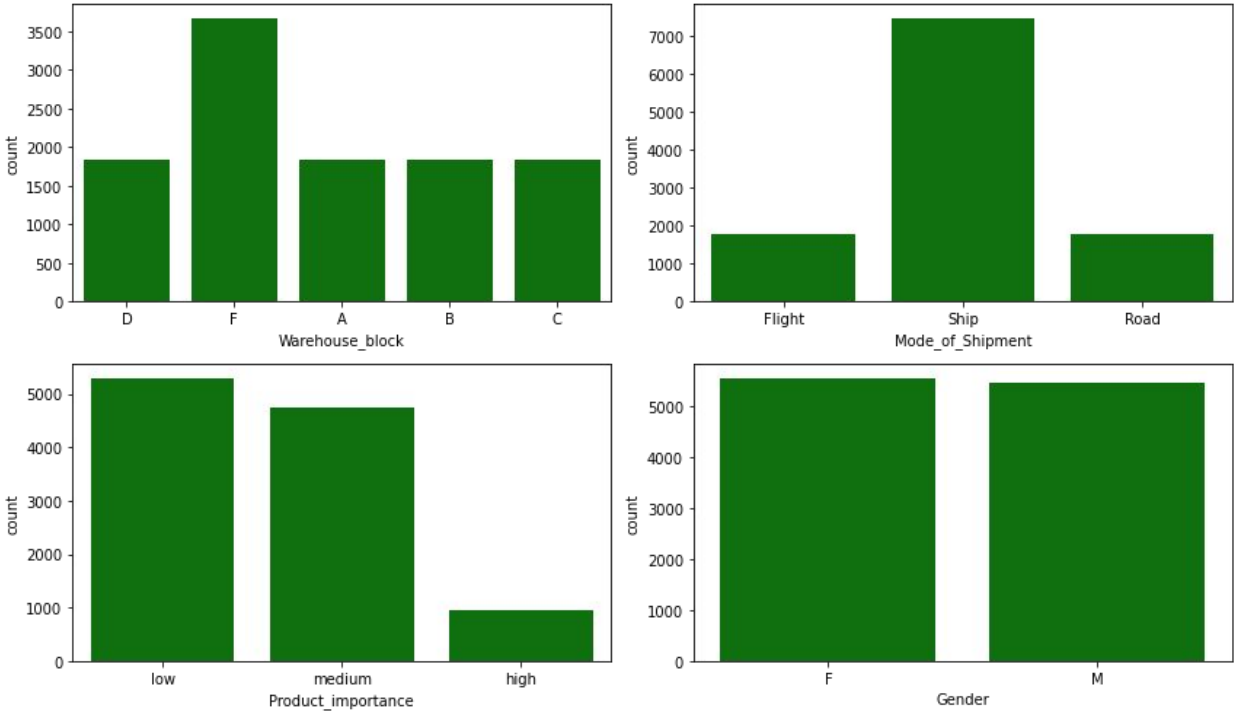


[2] Univariate Analysis: Data Numerik

1. Distribusi data numerik:

- `**Customer_care_calls**` mendekati normal namun cenderung skew ke kanan.
- `**Customer_rating**` memiliki jumlah yang mendekati setara untuk setiap nilai rating.
- `**Cost_of_the_product**` mendekati distribusi normal.
- `**Prior_purchases**` memiliki distribusi yang skew ke kanan dan terdapat outlier.
- `**Discount_offered**` terdapat terlalu banyak outlier, distribusi skew ke kanan.
- `**Weight_in_gms**` memiliki distribusi bimodal dan mean < median.
- `**Reached.on.Time_Y.N**` memiliki nilai 1 (terlambat) yang lebih banyak dari 0.

[2] Univariate Analysis: Data Kategorikal



Distribusi data kategorik:

- "Ship" di `**Mode_of_Shipment**` terlalu banyak dibandingkan "Flight" dan "Road".
- "F" di `**Warehouse_block**` terlalu banyak dibandingkan "A", "B", "C", dan "D".
- "High" di `**Product_importance**` terlalu sedikit dibandingkan "low" dan "medium".
- Distribusi `**Gender**` cukup seimbang dengan jumlah "F" yang lebih besar dari "M".

Rekomendasi untuk Pre-processing

Kolom numerik:

- Data skew kanan dapat ditransformasi logaritmik mendekati distribusi normal, lalu dapat dibuang outliernya (0.3% atau 5% terluar)
- Data terlalu banyak outlier: dibuang atau dijadikan max value

Dapat dikombinasikan:

- Value “high” dengan “medium” di **Product_importance** menjadi "medium-high"
- Value “Flight” dengan “Road” di **Mode_of_Shipment** menjadi "flight-road"

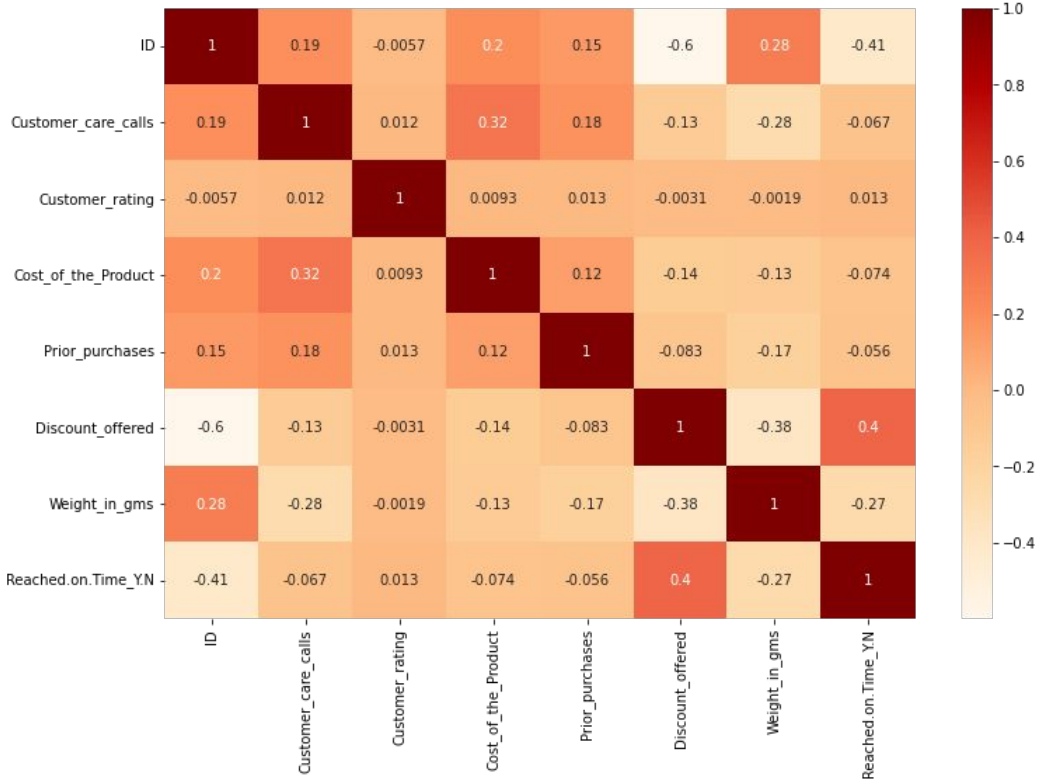
Dapat diedit: “F” di **Warehouse_block** diganti ke “E”

Dapat dibuang: **ID**

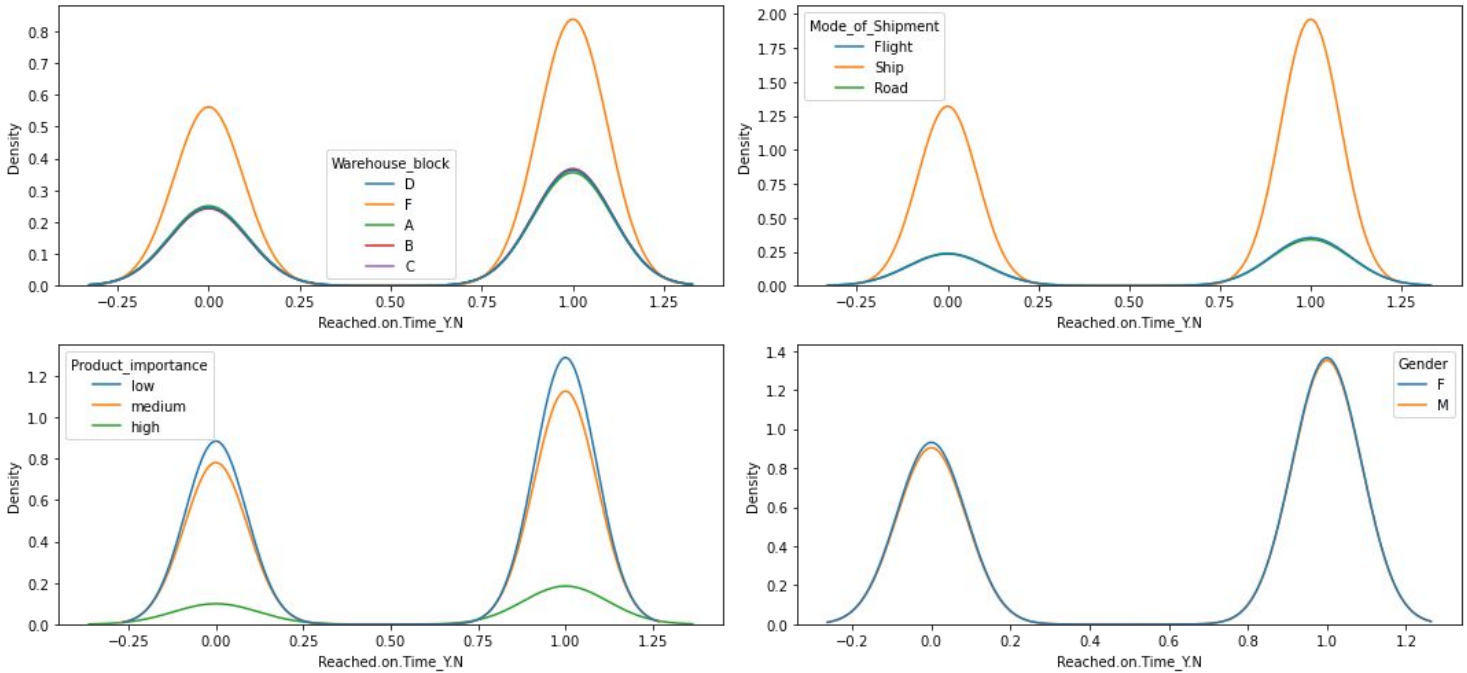
Dapat dilakukan Label Encoding: **Product_importance** dan **Gender**

Dapat dilakukan One Hot Encoding: **Warehouse_block** dan **Mode_of_Shipment**

[3] Multivariate Analysis: Data Numerik



[3] Multivariate Analysis: Data Kategorikal



[3] Multivariate Analysis

Bagaimana korelasi antara masing-masing feature dan label. Kira-kira feature mana saja yang paling relevan dan harus dipertahankan?

|R| < 0.05 :
Customer_rating

0.05 < |R| < 0.1 :
Customer_care_calls, **Cost_of_the_Product**, **Prior_purchases**

|R| > 0.1 :
ID, **Discount_offered**, **Weight_in_gms**

Feature `**Discount_offered**` cukup berkorelasi dengan `**Reached.on.Time_Y.N**` dengan nilai korelasi **0,4**.
Karena feature yang ada sudah sedikit, semua feature akan dipertahankan kecuali feature **ID**.

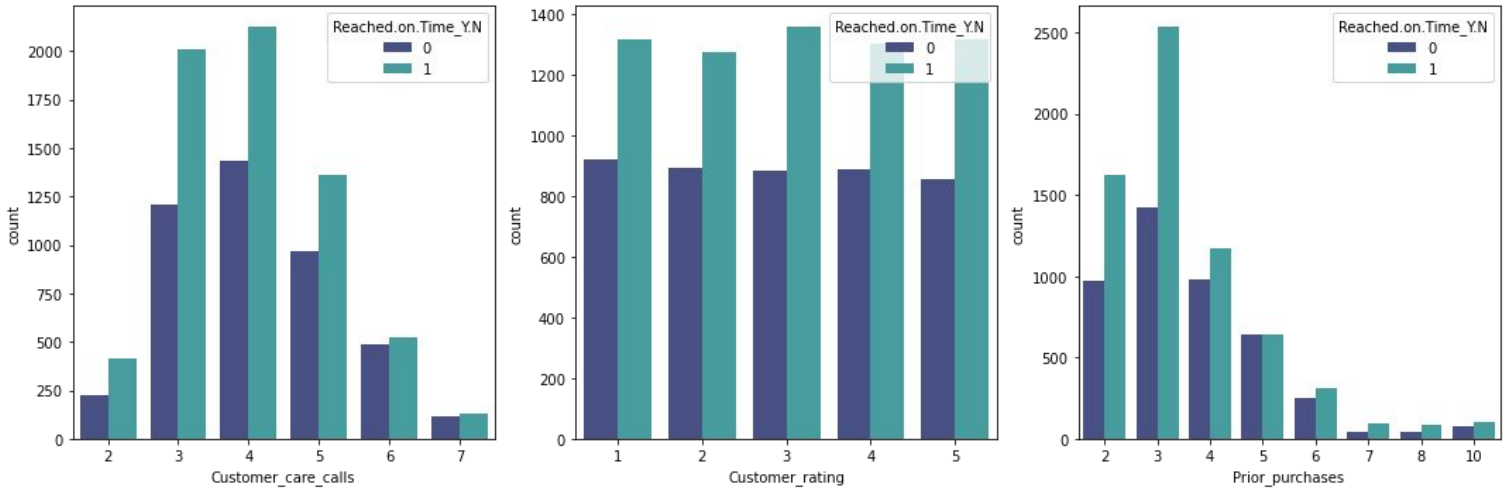
R = korelasi feature terhadap label

Bagaimana korelasi antar-feature, apakah ada pola yang menarik? Apa yang perlu dilakukan terhadap feature itu?

Pola menarik dapat terlihat di antara kolom **Discount_offered** dan **Weight_in_gms** dengan korelasi -0,38.

Kolom **Weight_in_gms** dapat didrop karena korelasi terhadap label lebih kecil dibandingkan **Discount_offered**, namun karena jumlah feature yang sudah sedikit, ada kemungkinan kedua kolom tersebut tetap dipertahankan.

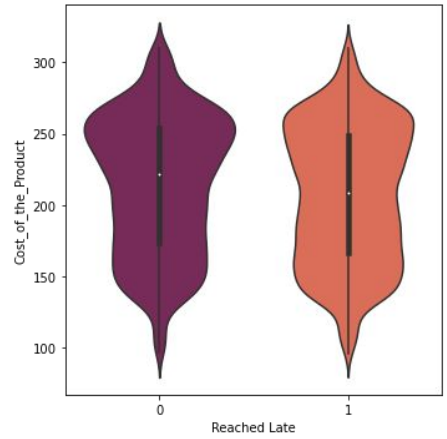
[4.1] Business Insights: Feature vs Label (1)

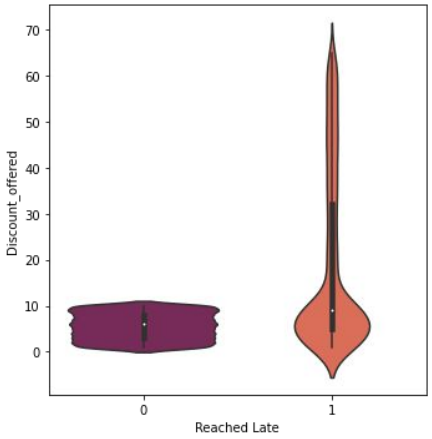


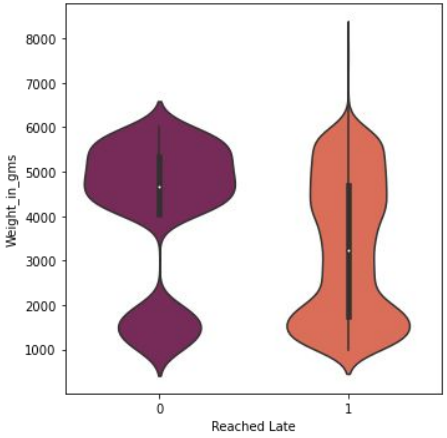
- `**Customer_care_calls**`: Mayoritas pelanggan akan menelepon CS sebanyak 3-4x terlepas produk terlambat atau tidak.
- `**Customer_rating**`: Pelanggan dengan produk tepat waktu paling banyak memberikan rating 1 dan pelanggan terlambat paling banyak memberikan rating 3.
- `**Prior_purchases**`: Mayoritas pelanggan telah melakukan pembelian sebanyak 3x sebelumnya, terlepas produknya terlambat atau tidak.

[4.1] Business Insights: Feature vs Label (2)

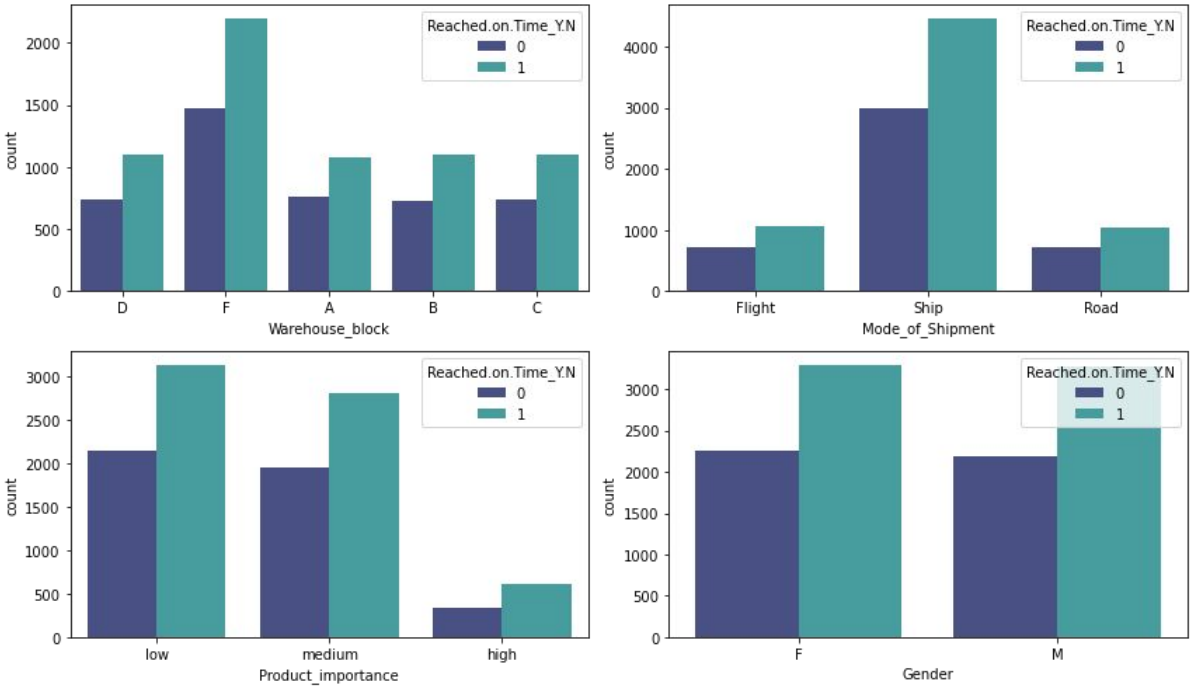
- `**Cost_of_the_Product**`: Produk yang terlambat memiliki median harga yang lebih rendah dibandingkan dengan produk yang tepat waktu.
- `**Discount_offered**`: Produk yang diberikan diskon besar cenderung datang terlambat dan produk yang diberikan diskon < $10 cenderung datang tepat waktu.
- `**Weight_in_gms**`: Produk yang memiliki berat 4000-5500 gram cenderung datang tepat waktu, sedangkan produk yang datang terlambat cenderung memiliki berat yang lebih ringan. IQR berat produk tepat waktu lebih sempit dibandingkan IQR berat produk terlambat.





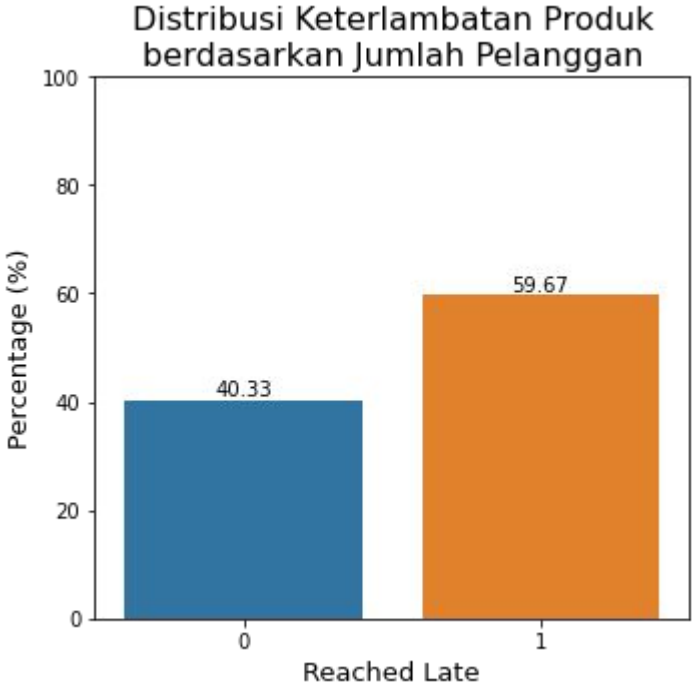


[4.1] Business Insights: Feature vs Label (3)



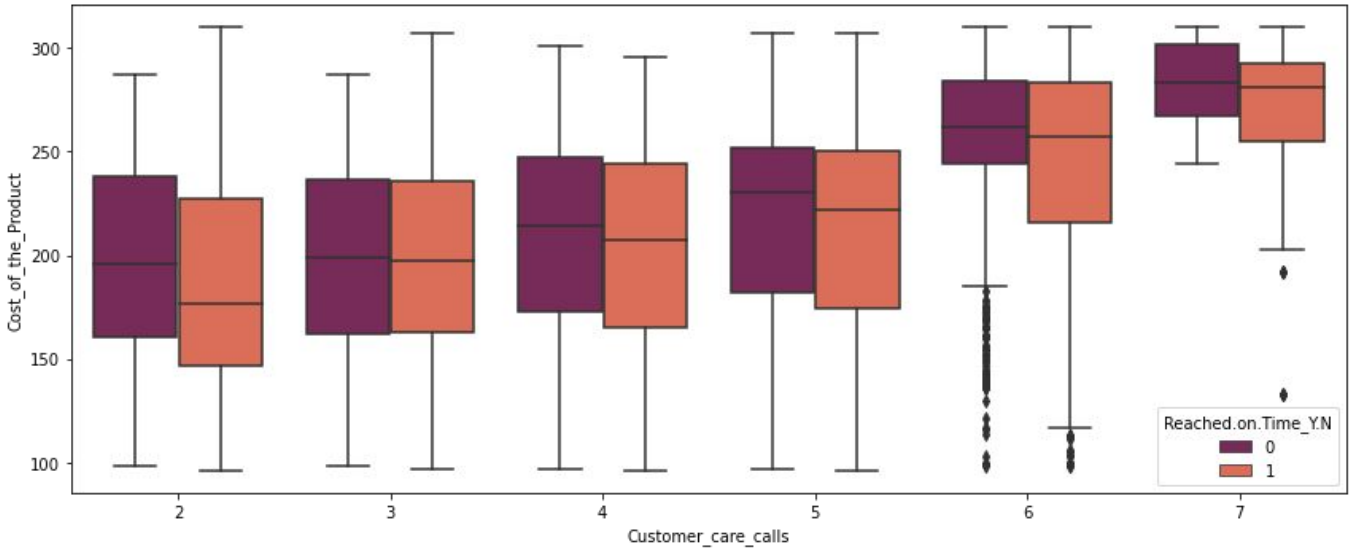
- **`Warehouse_block`**: Mayoritas produk dikirim dari Blok F, terlepas produk datang terlambat atau tidak.
- **`Mode_of_Shipment`**: Mayoritas produk dikirim menggunakan kapal, terlepas produk datang terlambat atau tidak.
- **`Product_importance`**: Mayoritas produk memiliki prioritas rendah, terlepas produk datang terlambat atau tidak.
- **`Gender`**: Mayoritas produk datang terlambat, terlepas dari jenis kelamin pelanggan.

[4.2] Business Insights: Keterlambatan Paket berdasarkan Jumlah Pelanggan



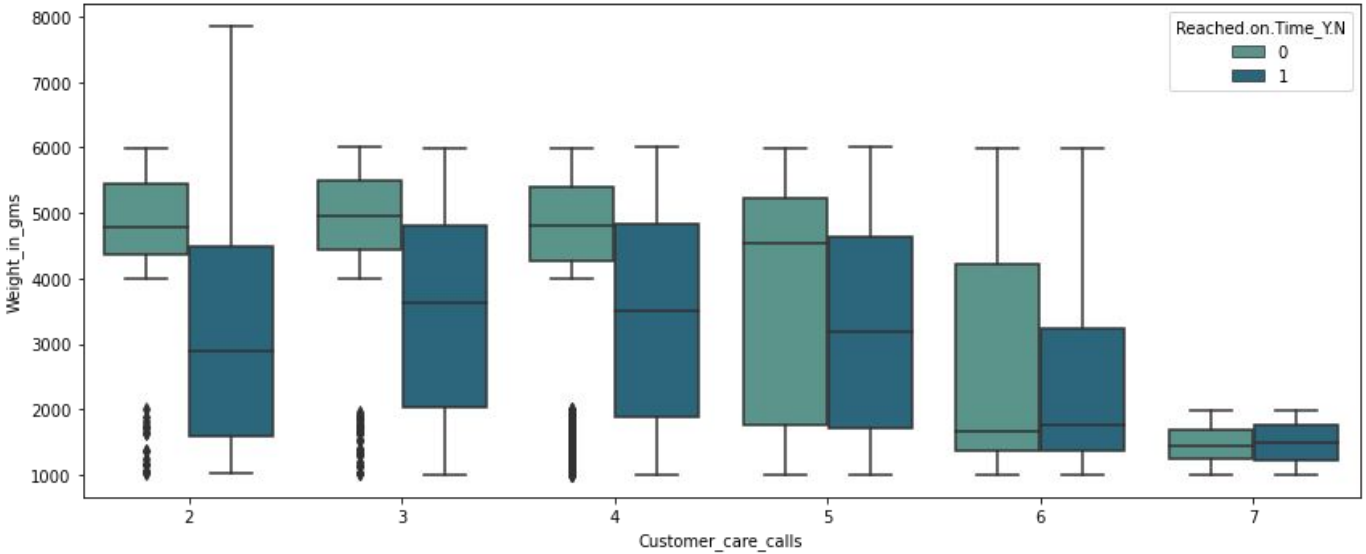
Dari 10999 data pengiriman produk, terdapat 59.67% produk yang datang terlambat dan 40.33% produk yang datang tepat waktu.

[4.3] Business Insights: Calls vs Product Cost



Semakin tinggi harga sebuah produk, pelanggan akan semakin sering menelepon CS.

[4.4] Business Insights: Calls vs Weight



Semakin rendah berat sebuah produk, pelanggan akan semakin sering menelepon CS.